EOSC-SYNERGY

EU DELIVERABLE: D4.1

Best Practices Elicitation including Data Management Plans

Document Identifier:	EOSC-SYNERGY-D4.1
Date:	29/2/2020
Activity:	WP4
Lead Partner:	UPV
Document Status:	APPROVED
Dissemination Level:	PUBLIC
Document Link:	https://u.i3m.upv.es/dirih

Abstract:

This document describes the thematic services that expose the applications and data to the scientific community. This is the first document analysing the thematic services and serves as a guideline for understanding their needs, commonalities and particularities. The 10 thematic services of EOSC-SYNERGY are reasonably different in maturity, requirements, technological needs and approach, which guarantees that the expansion of the EOSC capacity addresses multiple dimensions and challenges. The document includes the expected verification means and the identified Data Management Plans.

EOSC-SYNERGY receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 857647

I. Copyright Notice

II. Delivery Slip

	Name	Partner/Activity	Date
From	Ignacio Blanquer	UPV/WP4	18/2/2020
Reviewed by	**Moderator:** Jorge Gomes **Reviewers:** Mario David	LIP/WP1 LIP/WP3	21/02/2020 24/02/2020
Approved by	PMB	PO	24/02/2020

III. Document Log

Issue	Date	Comment	Author/Partner
v1	02/12/1999	TOC and initial draft version	I.Blanquer / UPV
v2	15/01/2020	Description of the Thematic Services	A. Calatrava / UPV, A. Azevedo / LNEC, J. Sánchez / INDRA, T. Kerzenmacher / KIT-IMK, J. Astalos / IISAS, M. Dobrucky / IISAS, S. Capella / BSC, L. del Caño / CNB, A. Krenek / CESNET, A. Rubio / CIEMAT, F. Benincasa / BSC
v3	27/01/2020	Resources and technical services	I. Blanquer / UPV
v4	5/02/2020	DMPs	A. Calatrava / UPV, A. Azevedo / LNEC, J. Sánchez / INDRA, T. Kerzenmacher / KIT-IMK, J. Astalos / IISAS, M. Dobrucky / IISAS, S. Capella / BSC, L. del Caño / CNB, A. Krenek / CESNET, A. Rubio / CIEMAT, F. Benincasa / BSC
v5	14/02/2020	First Draft for internal review	I. Blanquer / UPV
v6	24/02/2020	Final Document	I. Blanquer / UPV

IV. List of Acronyms

Acronym	Description
AAI	Authentication and Authorisation Infrastructure
AEMET	Spanish State Meteorological Agency
AERONET	AErosol RObotic NETwork
B2FIND	EuDat Discovery service based on Metadata
B2SAFE	EuDat Service for distributing and storing large volumes of data
B2STAGE	EuDat service for data ingestion
CCMI	Chemistry-Climate Model Initiative
CEDA	Natural Environment Research Council's Data Repository for Atmospheric Science and Earth Observation
CF	Climate and Forecast
CORSIKA	COsmic Ray SImulations for KAscade
CS	Consortium Spatial Information
CSW	Catalog Service Web
DEM	Digital Elevation Model
DIRAC4EGI	Distributed Infrastructure with Remote Agent Control for the European Grid Initiative
DMP	Data Management Plans
DOI	Digital Object Identifier
DREAM	Dialogue on Reverse Engineering Assessment and Methods
DYNAFED	Dynamic Federations system
ebRIM	Registry Information Model
EC3	Elastic Compute Clusters in the Cloud
EGI	European Grid Initiative
EIRENE	European Environmental Exposure Assessment Network
ELIXIR	Life Sciences ESFRI
EMODNET	The European Marine Observation and Data Network
EMPIAR	Electron Microscopy Public Image Archive
EOSC	European Open Science Cloud
EPA	Environmental Protection Agency's

EPANET	Water distribution system modeling software package from the United States EPA
ERIC	European Research Infrastructure Consortium
ESFRI	European Strategy Forum on Research and Innovation
EuDat	Collaborative Data Infrastructure for Data Preservation
FAIR	Findable, Accessible, Interoperable, Reusable
G-CORE	Earth observation data processing software from INDRA
GA4GH	Global Alliance for Genomics and Health
GEANT4	Toolkit for the simulation of the passage of particles through matter
GEE	Google Earth Engine
GPU	Graphics Processing Unit
HDF	Hierarchical Data Format
I2PC	Instruct Image Processing Center
IdP	Identity Providers
IGAC	International Global Atmopsheric Chemistry
IM	Infrastructure Manager
INGENIO	Spanish Earth Observation Satellite
INSTRUCT	Integrated Structural Biology Infrastructure
JSON	JavaScript Object Notation
LAGO	Latin American Giant Observatory
LANDSAT	Earth Resources Technology Satellite
LSDF	Large Scale Data Facility
LSDMA	Large-Scale Data Management and Analysis
MODIS	Moderate Resolution Imaging Spectroradiometer
MSWSS	Modelling Service for Water Supply System
NAMEE	Northern Africa, Middle East and Europe
NASA	National Aeronautics and Space Administration
NCEI	National Centers for Environment Information
netCDF	Network Common Data Form
netCDF	Network Common Data Form
NWP	Numerical Weather Prediction
O3AS	Ozone (O3) Assessment
OGC	Open Geospatial Consortium
OGC SOS	Open Geospatial Consortium Sensor Observation Service

OneData	Distributed Data Management solution from Cyfronet
OPENCoastS	Coastal circulation on-demand forecast
OpenEBench	Benchmarking service for Bioinformatics from ELIXIR
PAZ	Spanish Earth observation and reconnaissance satellite
POSIX	Portable Operating System Interface for X
QFO	Quest for Orthologs
RECETOX	Research Centre for Toxic Compounds in the Environment at Masaryk University
ROOT	Data Analysis Framework from CERN
SAPS	Serviço Automático de Processamento do SEBAL
Scipion	Cryo em image processing framework. Integration, traceability and analysis
SDS-WAS	Sand and Dust Storms Warning Advisory and Assessment System
SEBAL	Surface Energy Balance Algorithm for Land
SGE	Sun Grid Engine
SIC	Satellite Imaging Corporation
SMOS	Soil Moisture Ocean Salinity
SPARC	Stratosphere-troposphere Processes and their Role in Climate
TCGA	Cancer Genome Atlas
UAV	Unmanned Aerial Vehicles
UMSA	Untargeted Mass-spectrometry Analysis
WCD	water-Cherenkov detectors
WebDav	Web Distributed Authoring and Versioning
WMO	World Meteorological Organisation
WORSICA	Water mOnitoRing SentInel Cloud plAtform
ZBGIS	Basic Slovak database for GIS
Zenodo	OpenAIRE repository for Open Science

Table of Contents

Executive Summary

EOSC-SYNERGY aims at expanding the uptake of EOSC by building capacities. Thematic services constitute an important part of EOSC-SYNERGY and are the final layer that is exposed to final users. Therefore, the expansion of the capacity of the thematic services will require improved platform services and improved infrastructure services.

EOSC-SYNERGY has identified ten thematic services addressing four scientific areas (Earth Observation, Environment, Biomedicine and Astrophysics). Those thematic services are heterogeneous, addressing a wider range of requirements, maturity level, user targets and usage models. In the area of Earth Observation, services address the monitoring of coastal changes and inundations, the processing of satellite image data and the estimation of forest mass, addressing different types of targets. In Environment, the thematic service covers the monitoring and protection of ozone, the forecast of sand and dust storms, the simulation of water network distribution and untargeted mass-spectrometry analysis for toxics. In Astrophysics, the project aims at setting up an European service for the Latin American Giant Observatory and in biomedicine EOSC-SYNERGY covers the benchmarking of Genomic data processing tools and the processing of Cyron-electron microscopy imaging.

In the frame of EOSC-SYNERGY, these thematic services will improve in terms of authentication and authorisation, resource management, job scheduling, data management and accounting. Not all the services have identified gaps in all the previous aspects so each thematic service will focus the adaptation in the aspects that are more relevant according to the bottlenecks.

In a preliminary analysis performed by all thematic services several technical commonalities and differences have been identified. All thematic services share the importance of using a robust Authentication and Authorisation Infrastructure (AAI) compatible with the ones used by the target institutions. EGI Check-in has revealed to be a widely accepted choice. With respect to resource management, all services have the interest of dynamically provisioning processing resources, most of the cases on demand. Infrastructure Manager (IM) and the Elastic Compute Clusters in the Cloud (EC3) client have been identified by most of the thematic services as a technology capable of filling in this gap. Regarding job Management, most thematic services use batch queues, which could be extended to support containerised jobs. The usage of Kubernetes to orchestrate microservices and job queues of containers is also considered. The most challenging part is the management of data. Thematic services have identified important issues on transferring and accessing large volumes of data and require smart caching, advanced data transferring and massive persistent data storage. Solutions available in the EOSC marketplace will be studied and prototyped before adapting them into the thematic services. Finally, monitoring will be inherent to the usage of platform services.

The thematic services expect to reach a workload between 400 and 46.500 CPU hours per week (an accumulated 71K CPU hours per week) consumed by up to 10k jobs per week requiring a median of 16 GB RAM and 15 GB of storage per job. The persistent storage ranges from 2 GB to 500 GB (a median of 100GB and a total of 1 PB). This workload is not straightforward and it will require involving additional resource providers.

The thematic services have also defined a set of performance metrics grouped into five categories (impact on users, on Capacity and Capability of the service, on Scientific Outreach, on the usability of the

service and on Cross-Fertilization). These metrics can provide quantitative indicators of the performance of the thematic services and how they improve.

The last analysis of the Thematic Services at PM6 also defined the Data Management Plans (DMPs) for the 10 Thematic Services. These DMPs will be improved progressively as the Thematic Services evolve during the project.

Thematic services constitute a key activity to evaluate the impact of the capabilities in EOSC-SYNERGY with respect to adopting mature and scalable services, software and service quality assurance, increased resource capacity and improved user skills.

1. Introduction

EOSC-SYNERGY is a collaborative project involving several institutions from different European countries working together to combine their knowledge and expertise to expand the capability and capacity of EOSC. One of the key activities of EOSC-SYNERGY is to improve and expose a number of thematic services identified at the proposal writing time. These services are described and analysed in this document, which goes through their needs, current and planned architecture, expected impact and Data Management Plans.

This report belongs to the WP4, “Capacity building for Thematic Services”, and will be key for the selection of the technologies in other work packages and to evaluate and measure the service performance.

1.1. Scope of the document

This report covers a summarised description of the ten thematic services of EOSC-SYNERGY. These services are grouped into four categories: Earth Observation, Environment, Biomedicine and Astrophysics. The analysis will lead to the identification of commonalities, best practices and common requirements, regardless of the thematic area of the service. The document will also set the basis for the evaluation of the improvement achieved at each case by defining metrics for the evaluation of the service performance.

1.2. Target Audience

The document is intended for both internal and external use. The main internal target of this document is the global team of technical experts of the EOSC-SYNERGY project, both infrastructure (WP2) and Service adoption (WP3), as well as the Workpackages for skills development (WP6) and dissemination (WP1). The document will also serve to guide external researchers interested in contributing to EOSC through their own services providing experiences and best practices.

Finally, this document will serve the evaluators of EOSC-SYNERGY to evaluate the progress of the action with respect to the metrics defined.

1.3. Structure of the document

This document is comprised of 6 sections and 1 appendix. After this introduction, Section 2 provides an overview of the thematic services of the project, covering a description of the service, the inventory of data sources, technical details for the solution, gaps and bottlenecks and an analysis of the resources needed. Section 3 covers the metrics and the baseline for the validation of the project. Section 4 identifies several canonical application architectures coming from the study of the different services and section 5 includes the Data Management Plans. Finally, section 6 covers the conclusions and appendix A includes a glossary of the terms used in the document.

2. Description of the Thematic Services

This section describes the thematic services with a subsection per thematic service. A detailed description of each thematic service is provided as an annex. In this section we include a description of the thematic service including the data sources required for its operation and the gaps and bottlenecks. A joint analysis of the technology and resource demands is provided in the next section.

There are four thematic areas in EOSC-SYNERGY, represented in figure 1. These four thematic areas (Earth Observation, Biomedicine, Environment and Astrophysics) comprise a total of 10 thematic services. A separate subsection for each service is provided in this section.

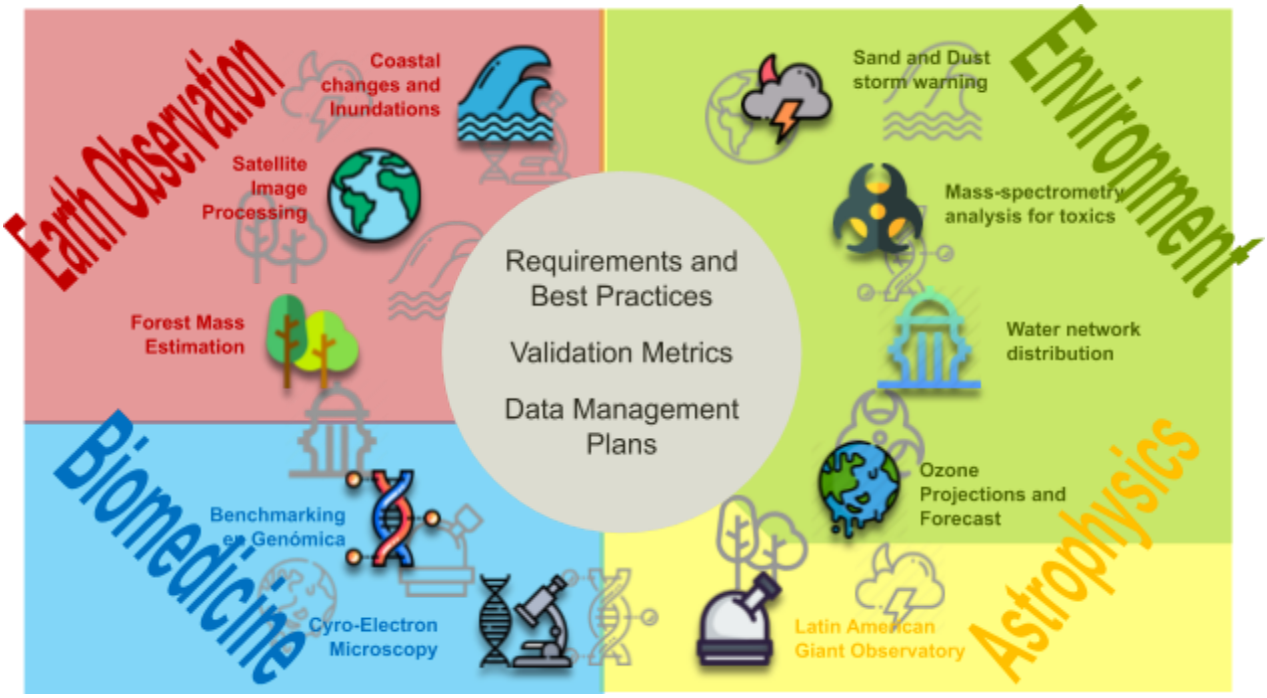

Figure 1: Thematic services by areas.

2.1. WorSiCa - Water mOnitoRing SentInel Cloud plAtform

2.1.1. Description of the Thematic Service

Water mOnitoRing SentInel Cloud platform (WorSiCa) is a service that detects the coastline, coastal inundation areas and the limits of inland water bodies using remote sensing (satellite and Unmanned Aerial Vehicles - UAVs) and in situ data (from field surveys). This thematic service aims at integrating multiple-source remote sensing and in-situ data to determine the presence of water in coastal and inland areas. It is applicable to a range of purposes, from the determination of flooded areas (from rainfall, storms, hurricanes or tsunamis) to the detection of large water leaks in major water distribution networks. It builds on components developed in both national and European projects, integrated to provide a one-stop-shop service for remote sensing information, integrating data from both the

Copernicus satellite and drone/unmanned aerial vehicles, validated by existing online in-situ data. The WorSiCa service will be available without cost to all european public research groups. The private sector will be able to use the service, but some usage costs may be applied, dependending on the type of resources needed by each application/user.

The integration of the WorSiCa service in the EOSC infrastructure will boost the usage of the service at an European level. This service will enable the research communities to generate maps of water presence and water delimitation lines in coastal and inland regions. These products can be useful for emergency and planning methodologies in case of inundations or reservoir leaks. In particular, the service promotes 1) the preservation of lives during an emergency, supporting emergency rescue operations of people in dangerously inundated areas, and 2) the efficient management of water resources targeting water saving in drought-prone areas.

The following impacts and benefits are expected from the WorSiCa service:

1. Fostering the use of the service thanks to the dissemination provided through EOSC channels and the availability of computational resources for its operation;
2. By using the EOSC infrastructure services, the costs in maintenance and acquisition of computational power are no longer attributed to the research communities;
3. The products delivered by the service will be widely used by the research communities and private companies in a panoply of distinct applications that can range from inundation to inland water bodies characterization, or even to extreme events such as rain flows or damn ruptures.

The service will also have an important impact on:

- Emergency: provide a fast access to inundated areas to support emergency rescue operations;
- Support management decisions on hydraulic infrastructures operation to minimize damage downstream;
- Climate change mitigation: minimize water losses and reduce water mains operation cost;
- Provide an early detection of water leakages in difficult-to-access water transportation networks, promoting their fast repair.

2.1.2. Data sources

The WorSiCa service operates on a diverse range of data. In order to determine the water masks, several types of images can be used, ranging from satellite to drone imagery. The satellite data are provided by: i) the Copernicus satellite and ii) the Pleiades satellite. The UAV data can be uploaded by the users. The algorithm to calculate the coastlines combines the usage of imagery data with the tidal information provided by the operational hydrodynamic prediction systems implemented in EOSC-hub OPENCoastS service and also by the EMODNET-Physics data. The portal EMODNET-Physics gives the WorSiCa service the complementary data to validate the results of the service with the tidal gauge data available for the European countries.

Data	Owner	Storage	How to access it

Copernicus data (ESA sentinel imagery)	ESA	Local servers[1]	Via the WorSiCa service account with ESA
Pleiades	Satellite Imaging Corporation (SIC)	Local servers	Using personal accounts
UAV	Users	Local servers	Uploaded by the users
Bathymetry data and Physics portals	EMODNET-Bathymetry	Local servers	Downloaded by WorSiCa
Sea surface height information	EOSC-OPENCoastS	Local servers	Connected to WorSiCa
Topography data	SRTM30 near-global digital elevation model (DEM) - U.S. Geological Survey	Local servers	Downloaded by WorSiCa

Table 1: Data Sources for the WorSiCa Service.

The EMODNET data (Bathymetry and Physics portals), are freely available to the public and are downloaded by the WorSiCa service for each application. The EOSC-OPENCoastS service will be connected to the WorSiCa service to provide sea surface height information for each application. Each user will also be able to upload local data from field campaigns.

2.1.3. Gap and Bottlenecks analysis

The WorSiCa service uses large sets of imagery data to produce the water indexes masks, therefore the main bottlenecks of the service are the:

- Downloading satellite data from operational providers (e.g. ESA or Pleiades): providers have implemented limitations on the download speed and number of concurrent downloads each user can do;
- Storage of the images needed for the algorithm to calculate the water and vegetation indexes: two reasons for that are 1) due to the download limitations imposed by providers, satellite imagery must be stored 'temporarily' so they can be quickly used for processing, 2) satellite imagery is high-resolution (each Sentinel-2 image is between 800MB and 1.2GB, while Pleiades is Very High-Resolution with a larger size) and extra storage must be provided to store processed intermediate products;
- The computation resources, where the GPU and RAM are highly recommended to speedup the image processing and to prevent bottlenecks on using the service during processing.

Since the algorithm is operating directly on the images downloaded from the operational providers or uploaded by the users, the computation can be easily distributed and broadcasted to several machines, taking advantage of the independent nature of the processing algorithm for each block of the image. Therefore, a dynamic cloud infrastructure with GPU resources is the most appropriate configuration for the WorSiCa service.

[1] Each application of the WorSiCa service will download the needed data to local servers in order to optimize the computational access of the input data.

2.2. G-Core

2.2.1. Description of the Thematic Service

G-Core is a production-ready technology used as a service at ESA's and national programs led by INDRA for the acquisition, storage, cataloguing and processing data from several EOS missions. G-Core provides two main functionalities:

- A Data Manager for spatial and non-spatial purposes; Ground Control points, GPS data, DEM, meteorological data, etc.
- A Processing framework to host external processors developed by third parties to generate added value products based on Satellite imageries.

The objective of the adaptation of the thematic service is to explore the sustainability of the EOS services exposed through the creation of added-value products through the integration of G-Core as a data manager.

The G-Core service targets the following three user profiles:

- EO data for the science community to use the satellite data in the scientific studies.
- EO data for public organizations to use the satellite imageries as background data.
- EO data for value adders to create added value products from satellite images.

The expected impact of the adaptation of the service is to democratize the usage of EO data out of the scope of nominal fields. It will help to define new products and services mixing Earth Observation data with other types of data for scientific and social environments.

2.2.2. Data sources

G-Core has been applied to process data from different public and private missions, such as:

- Sentinel data (public mission of EU/ESA).
- SMOS (ESA Mission for Scientific community).
- PAZ mission (Spanish Earth Observation program. Radar Mission).
- INGENIO (Spanish Earth Observation program. Optical mission)

Sentinel, SMOS and Paz are operational missions and Ingenio is currently under development. Due to the different missions involved, the next table shows the maximum size of products per mission with the goal of having an idea of the total amount of data to be processed, archived and delivered per product within the frame of the different missions previously mentioned:

MISSIONS	Products (Max per product in MB)			
Sentinel-1	SAR L0	SAR L1 SLC	SAR L1 GRD	SAR L2 OCN
	4000	7680	2000	14
Sentinel-2	L1c	L2A		

	600	800		
Sentinel-3	OLCI L0	OLCI L1	OLCI L2	
	9500	28500	28400	
Sentinel-5p	UV	UVIS	NIR	SWIR
	5600	5700	5700	2600
SMOS	L0 Consolidated Measurement	L0 Correlated Noise Injection (long)	L0 Uncorrelated Noise	L0 HKTM
	57.88	25.28	1.96	1.23
	L1a Measurement Product	L1b Measurement Product	G matrix	J matrix
	214	116	7820	1330
	L1c Measurement Product	Browse Product	IGSG File	VTEC_C
	521.34	4.38	0.86	0.16
	L2 SM DAP	Raw ECMWF Atmospheric Model	Native LAI	AUX_DGGFLO
	191.62	48.9	1536	33
PAZ	SM-S L0-SAR	SC L0-SAR	SL-S L0-SAR	HS-S L0-SAR
	763	1842	690	637
	SM-S L1B (EEC_SE)	SC L1B (EEC_RE)	SL-S L1B (EEC_SE)	HS-S L1B (EEC_SE)
	7352	3126	2006	1897

Table 2: Data sources used in the G-Core Thematic Service.

There are some variables that impact on the volume of data, the sensor/s involved in the mission (radar or optical), the nature of the mission (Scientific or commercial) and the product format requested by the user. The analysis is not only focused on the final user but also in the activities to be done in the ground segment in order to process the data incoming from the satellites.

2.2.3. Gap and Bottlenecks analysis

The service currently is limited by the following bottlenecks:

- Limited access to data repository, remotely or locally, due to network bandwidth restrictions.
- Infrastructure resources for processing and reprocessing large data sets.
- Data delivery volume. Increasing size of file to be delivered to users.

The requirements identified are:

- Interoperable interfaces to access and discover data from different sources.
- Cloud Infrastructure to use resources on demand for processing and reprocessing huge data.
- Distributed catalogue to ease the data discovering.

2.3. SAPS: Serviço Automático de Processamento do SEBAL

2.3.1. Description of the Thematic Service

SAPS is a service to compute the Surface Energy Balance Algorithm for Land (SEBAL) and similar information for estimating the evolution of forest masses and crops targeted to researchers in Agriculture Engineering and Environment. SEBAL can be used to increase the knowledge on the impact of human and environmental actions on vegetations, leading better forest management and analysis of risks.

By the deployment of a federated site of SAPS in EOSC, we will be able to facilitate European scientists to exploit the evapotranspiration estimation services from remote sensing imagery.

SAPS uses a cloud offering as a back-end. A processing engine submits jobs on a federated infrastructure. The code is available in https://github.com/ufcg-lsd/saps-engine, and a video is available in https://www.youtube.com/watch?v=-x3shbRMHkI.

2.3.2. Data sources

SAPS uses containers to facilitate the deployment of customizable versions of evapotranspiration processing algorithms that are broken in a three-stage pipeline: input data download, input preprocessing, and evapotranspiration estimation. SAPS comes with a number of implementations of these stages. In particular, it provides two different versions of the input download stage that use different data sources. The reference input download implementation uses multiple data providers. Landsat imagery is downloaded from the GEE platform. Meteorological information provided by the National Centers for Environment Information (NCEI - ftp://ftp.ncdc.noaa.gov/pub/data/noaa/), and elevation data provided by the Consortium Spatial Information (CS - http://srtm.csi.cgiar.org/) are downloaded from mirror servers of these services managed by the Federal University of Campina Grande (UFCG). The alternative implementation works similarly to the reference implementation, but downloads Landsat imagery from the USGS (http://ers.cr.usgs.gov/) service, instead of GEE.

2.3.3. Gap and Bottlenecks analysis

Currently, the service is limited in terms of computing and storage resources and data access. If the service is exposed to European users, it will require larger-scale deployment, integrated AAI, standardisation of services and improved scalability.

We expect to fill these gaps by integrating a dynamic management of processing resources that could facilitate horizontal elasticity and to integrate a coherent Authentication and Authorisation Infrastructure.

2.4. OpenEBench

2.4.1. Description of the Thematic Service

OpenEBench is a platform to support technical monitoring and scientific benchmarking activities carried by Life Sciences Communities. It is designed following a three-level architecture to facilitate support and interactions to Life Sciences communities at any maturity stage. OpenEBench aims to serve different researchers profiles e.g., software developers aiming to identify relevant datasets and metrics to test their software performance; 2) researchers looking for a mechanism to take informed decision on the best tools and/or workflows for their scientific problem at hand; 3) scientific communities that are interested to use a platform for their benchmarking activities, and others stakeholders e.g. funding agencies and policy-makers, who are interested to understand the current state-of-the-art of a particular area in Life Sciences.

The OpenEBench data model (github.com/inab/benchmarking-data-model) is essential to organize benchmarking-related data generated by any Life Sciences Community. Data is bundled and deposited in services like Zenodo and EuDat where they receive a DOI. It also connects with ELIXIR Core Data Resources and Deposition databases to deposit and/or access data produced/needed by the Scientific Communities activities.

OpenEBench already uses ELIXIR AAI, which is intended to evolve together with other services e.g. GEANT; as Life Sciences AAI in the context of the cluster EOSC Life. We expect also to make use of some of EOSC Life outcomes e.g. MyExperiment 2.0 - recently renamed as WorkflowHub; to access richly annotated workflows as well as RO-Crate - as a mechanism to store and retrieve specific instances of any workflow.

The work in this thematic service will consist on integrating those services in the EOSC Portal by exposing and deploying the benchmarked analytical workflows as well as extending its capacity through best practices and additional services. As impact, we expect Life Science researchers will have up-to-date collections of analytical workflows, which can be deployed across heterogeneous systems, organized by scientific communities around specific topics. This use-case will also provide best practices for organizing communities around scientific benchmarking activities.

2.4.2. Data sources

Currently, OpenEBench uses the Reference Datasets of three challenges:

- DREAM (Dialogue on Reverse Engineering Assessment and Methods) with 100 entries.
- The Cancer Genome Atlas (TCGA) with 885 entries.
- The Quest for Orthologs (QfO) with 1,838 entries.

All this data comes from public bioinformatics databases.

2.4.3. Gap and Bottlenecks analysis

OpenEBench has three levels of operation: Querying the existing benchmark results from already established Life Sciences Communities running their own benchmarking activities (level 1), the submission of user-specific tools through a Virtual Research Environment for running against reference benchmark data on data challenges (level 2) and the execution on-demand of user tools on user data (level 3 - in development). To widely support level 2 and to support level 3 OpenEBench will need to be extended to work on heterogeneous systems.

Secondly, OpenEBench needs to efficiently store processed data and workflows in a FAIR manner in a platform that could provide persistence, provenance and reproducibility.

2.5. Scipion Cryo-Electron Microscopy Service

2.5.1. Description of the Thematic Service

Scipion is an image processing framework used to obtain 3D maps of macromolecular complexes using cryo Electron Microscopy, whose development was started in the Instruct Image Processing Center (I2PC) located at CNB-CSIC.

Scipion is a desktop application that can be installed on personal desktops as well as big servers or clusters. Even though installation and configuration is intended to be as easy as possible there are some specific parts that could be more complex, such as MPI setup or GPU configuration. Besides, for a standard cryo-EM processing common desktop machines are clearly insufficient in terms of computing capability and storage, which could be a problem for many scientists that might not have access to powerful servers or GPUs. To overcome this limitation ScipionCloud was developed, resulting in a full installation of Scipion both in public and private clouds, accessible as public ‘‘images” that include all needed cryoEM software and just requires a Web browser to work as if it was a local desktop. These images could be used to easily deploy instances in the EGI Federated Cloud using the EGI Cloud Compute service or in AWS through their console. Furthermore, in the context of the Westlife project an alternative option to the use of fixed images was developed, based on the cloud orchestration tool Cloudify and the configuration management tool Puppet to automatically deploy and configure ScipionCloud software on a single machine or cluster in the EGI Federated Cloud. The service has a web front-end that allows to initiate the deployment by providing some information in a kind of wizard, in a similar way as the EOSC Applications on demand service works.

By making this service available in the EOSC marketplace researchers coming from an Instruct facility CryoEM session can have their data and preprocessing project available on a ScipionCloud cluster powered by EOSC compute resources on the back-end. This means that scientists with minimal computational background (or compute resources of their own) can access the latest tools as well as powerful computational resources to continue their processing.

2.5.2. Data sources

At the CryoEM facility (microscope) raw data is acquired in the form of 'movies', a typical session being around 1-2 TBs. In many facilities around the world they use Scipion to run an automatic preprocessing workflow in streaming mode, that helps to monitor the acquisition quality and also gives users a first idea on the structure they are going to obtain. Both raw data and the preprocessing project are given to users so they can later on continue processing using Scipion at their home labs but the aim is to send it to a remote storage such as OneData where the ScipionCloud service can access them later. Since movies are already processed in the facility a 'reduced-disk' option could be to copy only the preprocessing project with intermediate data (starting with micrographs), which would be of GBs order.

In the case of Instruct funded projects users can have their data for an established period of time after which it becomes public. Related to this and in the context of another European project, EOSC Life, we are working on a pilot where data and a rich workflow with the preprocessing steps are sent to the EBI EMPIAR database where data is kept private and then released after this embargo period but in principle it cannot be downloaded from there until it becomes public.

Data	Owner	Storage	How to access it
Microscope movies (1-2 TBs) and Scipion preprocessing project (~ 500 GBs). In principle only the project needs to be sent to the cloud.	Instruct (and users for an embargo period).	Users disk or OneData space.	Data is given to the user but could also be uploaded to a OneData provider at the facility. When launching the service data will have to be either uploaded or referenced and will be automatically mounted on the server.

Table 3: Data sources used in the Scipion Thematic Service.

2.5.3. Gap and Bottlenecks analysis

The service is currently limited by the following bottlenecks:

- Cloud resources insufficient: A typical processing workflow is composed of heterogeneous steps in terms of processing resources, some of them requiring powerful GPUs, other big RAM and some a high number of CPUs. In order to optimize the use of cloud resources a Resource Management able to deal with this scenario would be needed. Also storage is a limit since a typical CryoEM session starts with 1-2 TB raw data that can grow to another extra TB for project and intermediate results. This can be reduced if original movies are not transferred and kept (only needed on the preprocessing workflow done at the facility) to the order of GBs (~500).
  Besides each service usage (project) should provide resources for at least two weeks.
- Data transfer performance: Due to the amount of data to be transferred (especially if movies are considered) a fast and reliable mechanism is needed (gridftp, exploring commercial solutions such as globus online or aspera).
- Distributed and shared file system: Scipion and the software packages underneath expects a POSIX file system. The chosen solution (OneData) has to support this.

2.6. Latin American Giant Observatory - LAGO

2.6.1. Description of the Thematic Service

The Latin American Giant Observatory (LAGO) is an extended cosmic ray observatory, currently composed of a network of ten water-Cherenkov detectors (WCD) spanning over different sites located at significantly different altitudes (from sea level up to more than 5,000m) and latitudes across Latin America. LAGO is targeted to scientists working on High Energy Physics, effects of cosmic radiation, space weather, volcanology, etc.

Due to their extreme locations, data coming from WCD must be safely stored in repositories. However these raw data are not directly usable, first it should be pre-processed to clean noise from measurements, and finally analysed to become publicly available to the scientific community after a small waiting period. These phases constitute the main computing workflow for the collaboration that should be automated.

Furthermore, the complete LAGO dataset not only refers to these direct measurements performed by detectors, but also to the simulation of different cosmic ray phenomena in some energy ranges of interest. Simulations are arbitrary run by scientists in their computing resources, and generate data equivalent to the workflow mentioned above that should also be stored in repositories in order to avoid computing them again and to allow further comparison with similar outputs previously generated.

On the other hand, both processing and simulation data are generated with commonly used High Energy Physics applications, mainly with CORSIKA (including unofficial and customised releases), as well as GEANT4 and other self-designed statistical codes for the data analysis. All of them are high-throughput oriented, being able to be run on any computing facility and are especially suitable for virtualised environments.

There are four types of data for the LAGO Collaboration: raw (L0), cleaned (L1), analyzed (L2, L3), and simulated, the latter amounting to three the outputs corresponding to different versions of the software. In order to be properly stored, wrappers normalise raw data as outputs and generate metadata following the Dublin Core schema[2]. Additionally, non-native metadata tags are included such as; latitude, longitude, and altitude of the (real or simulated) WCD, the compilation environment, software releases, etc. The Collaboration has established 1 hour of measurements or simulations as the minimum suitable data-set. Thus, the minimal self-contained unit for real data (raw, cleaned and analysed data types) is a file with its linked metadata that spans 1 hour of measurements. However, simulations split the calculations (typically into 60 runs, one per minute) and the data-set contains the fabricated inputs and outputs with up to 180 data-set files including metadata.

This thematic service pursues the integration of the whole LAGO computing workflow in the EOSC portal, including data acquisition, curation, and long term visibility of their related results as well as the corresponding generated processing and simulation data. For this purpose, EOSC should not only

[2] https://www.dublincore.org/schemas/

provide computing resources or storage, but primarily services to allow harvesting data while maintaining long term preservation for reproducibility purposes, i.e. fulfilling FAIR principles.

Current status with respect to EOSC services: No EOSC services are yet integrated, though data curation processes and scientific codes execution methodologies have been implemented following *de facto* standards promoted by initiatives such as EGI.

2.6.2. Data sources

Typically, every WCD generates one measurement per hour producing ~ 200 MB files each, this is ~ 150 GB of raw data per month with a total of 720 files. For this reason, every file is considered the minimum data-set to be referenced and processed.They are classified as the Quality Level 0 (L0) of data through the Collaboration. Potentially, these files can originate 70-120GB of cleaned (L1) and 10-40GB of analysed data (L2 and L3). The one hour interval is kept as unit, resulting in 2160 files (between 2-160 MB each). The necessary amount of CPU resources to generate these files is small; around 35 mins on a Gold 6138 Intel core (2 GHz). However, all data-sets should be preserved for reprocessing if software errors can appear in the future. Furthermore, the classification of data levels is as follows:

- L0. Raw data.
- L1. Preliminary data: low resolution but the atmospheric pressure is corrected.
- L2: Ensures data quality to be used by experts from Astrophysics Community: fixed scalers by atmospheric parameters and the efficiency of the detector.
- L3. Ensures high quality to be used by researchers from other subjects or general public: the histograms are also corrected.

On the other hand, users can perform their own simulations, which should be re-usable and reproducible with other collaborators. A standard simulation using only CORSIKA (i.e. *background* simulation), results in a data-set of 4-10 GB (usually ~ 6 GB), but an *event* simulation could take 100 GB. In order to keep the 1 hour convention, both types of simulation are usually split into 60 runs, with an interval ranging from 15 mins to 10 hours, one task per minute. Thus, the complete workload of a *background* simulation is over 640 CPU/hours (Gold Intel core, 2 GHz). Additionally, to assure reproducibility, every input and output file of every run should be in the data-set comprising 180 files. (Note that additional files could complement the data-set if other applications were executed, but only the usual CORSIKA behaviour is considered in this Thematic Service).

Currently, there are 10 detectors installed (plus 11 planned), that can potentially transfer 18 TB/year of raw data to a centralised repository at UIS (Universidad Industrial de Santander, Colombia). However, this repository was not designed to be FAIR compliant, even public, hindering harvesting to process data in remote sites. As a consequence, members of the collaboration (~ 85 researchers) make rudimentary partial data-set copies to their local computing facilities, where they process the information they are interested in. In some cases, they upload their results and simulations.

The entire collaboration could generate up to 27 TB of raw, cleaned, and analysed data, plus 12-120 TB of simulated data in one year. Nevertheless, the availability of detectors is another issue, only 4 of them are actually achieving a 24/7 production operation. Moreover, an active user could submit 10 or 20 simulations per month, but actually researchers do not regularly submit simulations, and even some

may only run simulations sporadically. Therefore, a realistic estimation of the storage consumption of the LAGO Thematic Service could be around 3.6 TB/year of L(0-4) data corresponding to 4 WCDs and 2-8 TB/year corresponding to 25 active users.

Regarding the disclosure of the datasets, LAGO Collaboration requires a waiting period similar to the established ones for other HEP large experiments. Such a period should be set not only to properly exploit results by the Consortium prior to their public availability, but because the raw data must be pre-processed by the Consortium. Simulations will be made publicly available too, but the waiting period should be set by the user owning the data. Finally, full datasets will be made public to the Cosmic Ray, Radiation and Space Weather communities, complementing those provided by major consortiums such as the Pierre Auger Observatory. This data sharing will be largely improved by integrating it into the EOSC services. In this sense, it will be one advantage to add a-posteriori, additional information to datasets, in particular to link it to publications.

<table>
<tr><th>Data Type</th><th>Source</th><th>Owner</th><th>Visibility</th><th>Size</th><th>How to access it</th></tr>
<tr><td>Raw (L0)</td><td>Water-Cherenkov detector (WCD)</td><td rowspan="3">LAGO Collaboration</td><td rowspan="2">Private while analysed data are not available (public).</td><td>150GB/month (per WCD)</td><td rowspan="4">Username/Password into
- UIS repositories
- Partial mirroring in local data centers (for example at CIEMAT)</td></tr>
<tr><td>Cleaned (L1)</td><td>Raw data from WCD</td><td>70-120GB/month (per WCD)</td></tr>
<tr><td>Analysed (L2 and L3)</td><td>Cleaned data from WCD</td><td>Public after fixed waiting period</td><td>10-40GB/month (per WCD)</td></tr>
<tr><td>Simulated</td><td>1 User</td><td>User</td><td>Public after variable waiting period</td><td>Estimated:
1-4 sim. per month
(72-300 GB/year)
Max: 120GB/month</td></tr>
</table>

Table 4: Data sources used in the LAGO Thematic Service.

2.6.3. Gap and Bottlenecks analysis

Repositories are structured under some exploitation LAGO rules with data and metadata, which correspond to a four-layer scheme: private non curated, private curated, and public. Tools used for simulation are open source (mainly CORSIKA, also GEANT4, and ROOT).

Prior to the EOSC integration, the main drawback is the data preprocess needed in order to make the measured data meaningful. Service needs: Data storage, curation and harvesting; computing power for simulations.

2.7. Sand and Dust Storms Warning Advisory and Assessment System - SDS-WAS

2.7.1. Description of the Thematic Service

SDS-WAS is a World Meteorological Organisation (WMO) programme to improve capabilities for more reliable sand and dust storm (SDS) forecasts. The SDS-WAS mission is to enhance the ability of countries to deliver timely and quality sand and dust storm forecasts, observations, information and knowledge to users through an international partnership of research and operational communities. The users of this service come from the dust scientific community, including PhD students, researchers, meteo services, enterprises working on businesses affected by dust (solar, aviation, agriculture, etc)

The SDS-WAS, as an international framework linking institutions involved in SDS research, operations and delivery of services, addresses the following objectives:

- Provide user communities access to forecasts, observations and information of the SDS through regional centres connected to the WMO Information System (WIS) and the World Wide Web.
- Identify and improve SDS products through consultation with the operational and user communities.
- Enhance operational SDS forecasts through technology transfer from research.
- Improve forecasting and observation technology through coordinated international research and assessment.
- Build capacity of relevant countries to utilize SDS observations, forecasts and analysis products for meeting societal needs.
- Build bridges between SDS-WAS and other communities conducting aerosol related studies (air quality, biomass burning, etc).

The Regional Center for Northern Africa, Middle East and Europe (NA-ME-E) was created in 2010 in Barcelona, Spain, and it is jointly managed by the Spanish State Meteorological Agency (AEMET) and the Barcelona Supercomputing Center (BSC-CNS). Its web portal http://sds-was.aemet.es includes:

- In-situ and remote-sensing dust-relevant observations.
- Daily experimental dust forecasts from several organizations.
- Information and training material from several past workshops.
- News and events for the SDS-WAS community.

The SDS-WAS NA-ME-E Regional Center runs, collects and offers several numerical models outputs for dust forecast (e.g. https://sds-was.aemet.es/forecast-products/dust-forecasts). Model simulation (daily runs of 72 hours forecast, two variables) produce numerical outputs formatted in an international well documented standard (netCDF) and organized per year/month/daily files.

Through the integration of such services in the EOSC, a more complete set of derived services can be built and offered to a wider group of users. The geographical area of interest of this service is reaching less favoured countries and having the potential to increase the quality of life.

The integration of the numerical data of models simulations for dust forecast, observational data and data processing into the EOSC catalogue to disseminate data, improve FAIRness and share some data analysis results.

2.7.2. Data sources

All data is stored in an in-house shared storage file-system. Data can be classified in two types:

- Model outputs: a set of 12 NWP (Numerical Weather Prediction) model outputs of two variables (dust surface concentration at sea level and aerosol optical depth of the whole column) with 72 hours forecast (3/6 hourly) at various spatial resolutions from 0.33º to 0.5º approximately.
    - Two of these models are run in house in an HPC infrastructure.
    - The remaining are collected from partner institutions with a variety of protocols/methods: http, ftp, receiving, downloading, etc.
- Observations: to perform model evaluation and validate results, a set of observations is downloaded.
    - In situ: based on photometers worldwide network managed by NASA named AERONET.
    - Satellite: download of two different products of MODIS satellite provided by NASA.

Models outputs are processed to a common data standard following netCDF format and CF-1.6 conventions. Observations come in different formats, which are processed and formatted to be compared with model data.

2.7.3. Gap and Bottlenecks analysis

The service currently uses B2SAFE (for the storage and backup) and B2STAGE (for staging and moving data from HPC to Storage). However the service still has storage limitations. Along with the storage, the service has identified the following gaps:

- Lack of services needed for Data storage and curation as well as computing power for data analysis in the on-demand mode.
- Managing the lack of reliability of data sources, especially about observations (stations not available, not well calibrated, etc).
- Part of the data is not simulated locally but retrieved from partners servers, often in experimental mode, which led to gaps as simulations may have not been completed.
- Evaluation process of model vs observations (data analysis) off-line (nightly cron jobs).

2.8. UMSA: Untargeted Mass-spectrometry Analysis

2.8.1. Description of the Thematic Service

UMSA is an untargeted mass-spectrometry analysis service from RECETOX (Research Centre for Toxic Compounds in the Environment at Masaryk University) in the Czech Republic. The service is expected to evolve to a key component of the emerging EIRENE ESFRI. By means of the integration in EOSC, uniform access to data and computing resources will be provided, scaling the service to the target

European-wide user community. Typically, mass spectrometry is done in a targeted way to confirm or disprove the presence of a specific compound in a sample. On the contrary, we aim at processing data to correlating the whole spectra (ie. all the present compounds) with other data (social, medical, other sample analyses, etc.) to work with more complex hypotheses of environmental impacts on human health.

2.8.2. Data sources

Data are generated by mass spectrometers, typically tens of GB of raw data per sample. The use case will develop gradually from hundreds to tens of thousands of samples. The data are typically generated in vendor proprietary raw format. Therefore the acquisition process includes a well-defined protocol to convert the raw format to an open one (mzML), and to extract required metadata (origin of samples, conditions of the measurements etc). The resulting dataset is uploaded to the service data storage, and it is expected to be kept for long term (decades).

Initially, the data will be acquired at Recetox laboratories (Masaryk University, Brno, CZ), expanding to multiple labs participating in the EIRENE infrastructure.

2.8.3. Gap and Bottlenecks analysis

The data are unrecoverable, original samples cannot be re-acquired, therefore long-term data storage (even decades) is required, together with appropriate data curation. Although it may exceed the scope of EOSC, as project progresses, long-term sustainability will be evaluated on the Data Management Plans.
Tracking provenance of the secondary (derived) datasets, i.e. what was the exact process of generating them from the original source data, is fairly critical, as the results may differ dramatically with different settings. Galaxy provides an elementary framework for provenance tracking, and it must be interfaced to the community identifier tracking service.
Some of the tools were developed on low-resolution data. They are expected to produce correct, high-quality results. However, the current implementations run out of any feasible memory limits. Work on reimplementing the algorithms to deal with sparse data has already started.

2.9. MSWSS : Modelling Service for Water Supply System

2.9.1. Description of the Thematic Service

MSWSS is a service for analysis of water distribution networks with regards to the mitigation of hazardous events by the integration of existing on-line analysis of toxics in drinking water supply networks with water distribution network simulation (EPANET). Other potential uses of the service are rehabilitation planning and optimisation. Analysis of hazardous events (e.g. toxics propagation within pipe system) may be used for preparation of risk management plans for water utilities with potential to be extended to an on-line early warning system. In addition to the use by water infrastructure operators, the service could be used also for research and educational purposes.

Typically, the MSWSS service will be used periodically by the water infrastructure operator as needed (assumption is once per month) to compare outputs of simulation with measured data. Next use is for real-time monitoring and identification of hydraulic failures and risk analysis of hazardous events when the simulations will be submitted automatically by the MSWSS service. In addition to periodic processing the service will be used on-demand (e.g. for rehabilitation planning and optimisation). The most computationally intensive use will be failure (or water loss) analysis which will be based on the simulation of various scenarios (~ thousands of simulations running as one collection of jobs).
By the integration of MSWSS in EOSC it will be possible to offer the service to a wider community of potential users. The computing infrastructure of EOSC will enable modelling of more complex water supply systems and to increase the number of scenarios for the analysis. Scientific and educational communities will benefit from EOSC data sharing services.

2.9.2. Data sources

This section lists the main data sources required for the operation of MSWSS service, which will provide basic pre-processing tools needed for proper hydraulic simulation. Various open data sources can be used as complementary inputs for pre-processing and post-processing as well. Due to the character of input data, the main part of pre- and post- processing is provided by users themselves based on their need (actual status of MSWSS). Post-processing of outputs for large sets of simulations is planned to be run within EOSC.
Periodic run of service (monthly) is based on actual need of water supply operator for testing area (Bratislava city for example) and outputs are processed to check the difference between the outputs and SCADA data (actual measured data partly used as boundary conditions for hydraulic simulation and partly used for validation of results). There are several features (not all of them are used periodically) for post-processing of outputs depending on actual needs of infrastructure operator - hydraulic check of the system, risk analysis for hazardous events, hydraulic loss examination and preparation for rehabilitation planning. Based on the actual situation the post-processing is made usually manually on the user's side, but we plan to integrate it to MSWSS service.

Data Name	Owner	Storage	How to access it
GIS	Water infrastructure operator	local	Will be uploaded by owner
CIS	Water infrastructure operator	local	Will be uploaded by owner
SCADA	Water infrastructure operator	local	Will be uploaded by owner
DEM50	GKU	local	wms service
ZBGIS	GKU	local	wms service
openstreetmap	Openstreetmap.org	local	wms service

Table 5: Data sources used in the MSWSS Thematic Service.

2.9.3. Gap and Bottlenecks analysis

Due to the national legislation (e.g. in Slovakia) the operational data of water infrastructure operators can have confidential status. In that case the data protection measures will have to be implemented at all stages of data processing. The data will have to be stored only in MSWSS private storage and virtual machine protection measures will need to be negotiated individually with IaaS providers. However, the MSWSS service instance that will not need to process confidential data (e.g. for research or education purposes) could use EOSC computing and storage resources without restrictions.
Post-processing of job outputs will be enhanced for improving the representation and exploitation of data products. The current status of MSWSS is based on results from water supply networks for cities with around 300-400 thousands inhabitants. The service will need to be checked for a large network (around 10 million inhabitants) to check if the service will suffer from bottlenecks.

2.10. O3AS: Ozone (O3) Assessment

2.10.1. Description of the Thematic Service

O3AS will provide easy access to simulations of past and future stratospheric ozone levels. Stratospheric ozone protects life on Earth from harmful UV radiation. This protective stratospheric ozone layer has been attacked by anthropogenic chlorine containing substances (e.g. CFC-11 from foam production). To protect the ozone layer the Montreal Protocol (MP) and its amendments regulate / prohibit the use of ozone depleting substances. The MP requires a status assessment of stratospheric ozone every four years – the so-called "Scientific Assessment of Ozone Depletion"[3].

Commonly, numerous model simulations are performed to estimate past ozone decline and current and future rates of ozone recovery (e.g. as part of the chemistry-climate model initiative, CCMI). The simulations of the past require verification with observational data (e.g. using satellite instruments). The simulations of the future are used to calculate milestones relevant for policy makers, e.g. the time when ozone levels will be back to 1980 levels. Here, we propose a service to analyze ozone projections and calculate pre-defined milestones. An example of such a workflow is described in [4].

The service is of interest for users working on atmospheric composition, policy makers and interested citizens in general. For every Scientific Assessment of Ozone Depletion large data volumes (from different models and simulations) have to be analyzed to generate key metrics for policy makers (e.g. ozone return dates, see above). This analysis is always time critical and speeding up the turnaround will ease the assessment process for future cycles (e.g. 2022, the 2018 Assessment has been published in December 2018). In particular, robust results (including quantified uncertainties) of future ozone levels are required for impact studies to gauge potential damage.

[3] https://www.esrl.noaa.gov/csd/assessments/ozone/2018/
[4] https://www.atmos-chem-phys.net/18/8409/2018/

2.10.2. Data sources

Currently, publicly available data will be used from the Chemistry–Climate Model Initiative (CCMI) (stored at the CEDA Archive[5]). Simulated ozone data (3d+t) is available for the years 1960-2100: For example, monthly means of ozone will be used to estimate the temporal development of ozone, including a return date for e.g. 1980 (or other years). For such a task, the following estimate for expected data volumes applies:

The total amount of data for the proposed service will be around 300-500TB (140 years x 12 months x 20 models x ca. 8-15 GB per file), assuming low to medium resolution models. A future increase in model resolution will increase the amount of data significantly (factor 4 seems likely). The files are available in netCDF format. First, we will work with a local copy of the data to optimize the workflow (see below). Different access patterns of the local data will be tested to improve performance.

2.10.3. Gap and Bottlenecks analysis

The service identifies several bottlenecks mainly related to the data.

- Data availability (in particular of new simulations; a future data policy could be restrictive);
- Fast handling of big data (some standard tools generate large amounts of intermediate data during the (pre-) processing);
- storage of data (if data amounts increase significantly in the future, due to increased model resolutions or higher temporal sampling).

[5] http://data.ceda.ac.uk/badc/wcrp-ccmi/data/CCMI-1/output

3. Requirements of the Thematic services

Every thematic service has made an analysis of the technical services used for managing the users, computing and data. Considering the gaps and bottlenecks identified, this analysis also considers potential alternatives of the current technical services used to overcome such issues. The identification of technologies is being discussed within WP2 and WP3, which will issue recommendations according to the quality standards defined in WP3 and the availability of services and resources identified in WP2.

The first subsection describes the technical services by thematic service and the second subsection describes the workload requirements.

3.1. Technical requirements with respect to Services

The thematic services have to deal with five main technical requirements:

- Authentication and authorisation. Choosing the right authentication and authorisation means is key to reduce users' effort and reluctance (by adopting well known and trusted Identity Providers - IdPs) and provide a coherent mechanism for the authorisation among services. In the cases where the management of users' rights on data and resources is delegated to the underlying services (for example, to deploy resources with the credentials of the user or to define access control directly on the storage), special attention should be paid. Thematic services manage users on a portal and spawn resources and store data using centralised (application-manager) credentials may have more freedom. In the former case, most projects will rely on EGI-checkin using community or institutional IdPs.
- Workload management. Thematic services typically have to deal with the execution of batch jobs on execution queues to produce results to be exposed to the users. In this case, there are two alternatives: using a local queue on a set of dedicated resources or directly executing jobs in a shared execution pool. The former is addressed by using Slurm, Torque or custom services on top of Kubernetes. For the latter, HT Condor and DIRAC4EGI are the selected solutions.
- Resource Management. Thematic services require a cloud infrastructure to deal at least with part of the workload. Instantiation of resources can be directly performed on top of IaaS, but contextualization and configuration is typically requested to increase repeatability and matching resources to the existing workload. Infrastructure Manager (IM) and Elastic Compute Clusters in the Cloud (EC3) are the technologies mainly chosen by the thematic services.
- Data Management. This constitutes the main challenge of most of the thematic services, which demand an efficient way to store and access large-scale distributed data. In this regard, choices are open, and EOSC-SYNERGY has identified ONEDATA, DYNAFED+WebDav, EUDAT B2SAFE, or direct cloud Object Storage (S3 or Swift).
- Monitoring. Most of the thematic services do not have specific monitoring. Although this will be provided by the underlying infrastructure, applications will surely have to link to it to be able to recover in the case of failure.

The individual services identified (or planned) are described in tab_techservices. Row "Now" denotes the current solution used and row "Plan" indicates the solution that has been identified. In some cases,

there is no requirement on changing the current solution (single cell covering both rows) and in some cases there is a need for changing but no decision has been taken yet (indicated as TBD).

Service		WorSiCa	G-Core	SAPS	Scipion	LAGO	SDS-WAS	UMSA	MSWSS	O3AS	OpenEBench
AAI	Now	Local	user/pwd with SSO and Kerberos LDAP and CAS	Local authz token	Westlife AAI (AARC based).	Username and password	Web login with user & password	Life-science AAI	Local	None	ELIXIR AAI
	Plan	*EGI Check in*		*EGI Check in*	*EGI Check in*	*EGI Check in*	*Local*		*EGI Check in*	*EGI Check in*	Life Sciences AAI/EGI-Checkin
Workload Mng.	Now	Local batch system	GCore+ K8s	Own Scheduler	Torque	Cluster batch	Local	Galaxy/ Slurm	Local batch system	Cloudify	SGE
	Plan	HT Condor or DIRAC4EGI		K8s	Kubernetes	Cluster batch & EC3 / DIRAC4EGI	TBD	TBD	Slurm	IM / Other solution	GA4GH TRS/WES/ TES stack
Resource Mng.	Now	Manual	Rancher	Fogbow	Cloudify	Cluster batch	Local	Openstack	Manual	No specific need	OpenNebula
	Plan	IM, Ansible	Rancher + IM/EC3	IM/EC3	IM / Other solution.	Cluster & IM+EC3/VMops	TBD	TBD	IM, Ansible		
Data Management	Now	Local	ElasticSearch for the catalogue	OpenStack Swift	Local + OneData.	Local one	Local	Local filesys, Ceph	Local	LSDF (KIT-SCC)	Local
	Plan	Object stor. / EGI DataHub				B2 tools (B2FIND) & EGI DataHub	TBD	TBD	Local + OneData	WebDAV or OneData	Zenodo / EUDAT
Monitoring	Now	N/A	GCore	None	None	N/A	Nagios	Icinga (local)	N/A	N/A	UpTime robot
	Plan	ARGO		TBD	TBD	ARGO	TBD	TBD	TBD	TBD	TBD

Table 6: Technical services identified for each one of the thematic services. The services identified to fulfill the gaps are listed under the "plan" row.

3.2. Technical requirements with respect to Resources

The thematic services have also evaluated the amount of resources that will be needed for the service in production. This information is evaluated using historical records and estimations based on experiments. The requirements on resources are evaluated considering the different execution models, which may involve unplanned execution peaks, periodic runs, on-demand execution runs and combinations of them. Table 7 shows the requirements on computing by the different thematic services. The expected

CPU consumption is evaluated on CPU hours per week, number of jobs, memory requirements, storage requirement (input, output and temporal), and execution pattern.

The demand is heterogeneous, ranging from few CPU hours per week up to peaks on the order of tens of thousands of hours per week. Maximum values are marked in red, and minimum values on green.

	WorSiCa	G-core	SAPS	Scipion	LAGO	SDSWAS	UMSA	MSWSS	OpenEBench	O3AS
Total CPU Consumption by time (hours/week)	Max. 400 h/week	32 cores (60% usage)	200K CPU h/month	Larger than 0.4KCPU h/week	4,2K CPU h/week, 26 cores/day	878 CPU h/week	50-100 cores sustained, peaks of 1K CPU Cores	~20 CPU-hours per run	> 0.1K h/week	50-100 CPU cores -168 CPU h/week
Number of individual jobs by time (jobs/week)	2 jobs per user per week	100 jobs/ day for processing	4K jobs/month (single run)	~100 steps, some of them with hundreds of jobs.	2,4K jobs/week	1 daily job (248 CPUs x ~30 mins)	140 jobs/week, peaks of 1K jobs/week	on demand, ~20 jobs per application run	10 jobs/week, peaks >100	>100 jobs/week
RAM required by each single job.	4-16 GB	2GB to 16GB	16G bytes	up to 16 GB	up to 4GB	~2GB per core (max)	mostly few GB some datasets 100GB	< 1 GB	2GB to 8GB	16-20 GByte
Storage for each single job.	up to 550GB	0.001GB to 8GB	~4G bytes	<100 GBs	Usual:10 GB Max: 100GB	~1TB	Up to 100 GB.	~ 1 GB	150Mb	8-15 GByte per file
How the application is run	40-50 runs/month	Typically 4 runs/ day.	a batch of about 800 jobs.	Single large runs lasting for 1-2 weeks.	338 jobs/day	1 daily execution	Continuously & seldom for historic data.	1 / month per domain + var. demand	On demand and periodically.	Web-intf, triggers job runs
Size of the input data per job	< ~550GB	0.1MB to 2000MB	0.4 GB	<100 GBs	Up to 200MB	~50MB	Up to 100 GB	~ 800 MB	~70MB	16-20 GByte
Size of the Output data per job:	< ~10GB	0.1MB to 8000MB	~4 GB	<100 GBs	Up to 2GB	~900MB	Up to 100 GB	~ 200 MB	100MB	8-10 GByte

Table 7: Workload analysis with respect to the consumption of resources.

The analysis of the resources will be matched with respect to the inventory of resources in EOSC-SYNERGY. It is envisaged that additional resources should be accessed, so EOSC-SYNERGY will check with external distributed computing and storage infrastructure, sister projects and international collaborations to fill this gap.

	Min	Max	Total	Q1	Median	Q3
Total CPU Consumption (hours/week)	400 h/w	46,500 h/w	70,906 h/w	400 h/w	878 h/w	4200 h/w
Number of single jobs (jobs/week)	1 j/w	10,000 j/w	13,452 j/w	20 j/w	100 j/w	2400 j/w
RAM required by each single job.	<1 GB	100 GB	N/A	2 GB	16 GB	16 GB
Storage required by each single job.	0.001 GB	1 TB	1,788 GB	8 GB	15 GB	100 GB

Table 8 : Summary of the computing resources requested (minimum, maximum, total and quartiles).

Table 8 summarises the resource demand. The most representative values are the quartile thresholds. Therefore, we have to provide VMs of at least 16GB of RAM each to be able to fulfill 75% of the use cases demand. In the same way, with 15 GB of scratch space per job we will be able to deal with 50% of the thematic service requirements. Same reasoning can be applied to the CPU hours and the job thresholds. In this way, we could evaluate the extra cost of specific solutions. For example, the job throughput for half of the thematic services is of 100 jobs per week, which is a weak requirement with respect to current production schedulers. Reaching the 10K jobs per week may be challenging, but this just applies to one single case which may have a specific solution or may be worthwhile to apply some optimisation to the service (if feasible).

A similar analysis has been made to the storage. The result of the different thematic services is shown in table 9.

	WorSiCa	G-core	SAPS	Scipion	LAGO	SDSWAS	UMSA	MSWSS	O3AS	OpeneBench
Permanent storage required	20 TB	E.g. PAZ 36 TB	4GB per job	100GBS	5,6 -11,6 TB/ Year	878 CPU hours/week	100-200 TB	2 TB	300-500 TB	100 Gb.
Granularity (number of individual files)	For each job Small: 10 Large: ~2000 files	1K-100K	100 file per job	Single large experiments lasting for 1-2 weeks.	157K files/year 50KB to 2GB	1 daily job (248 CPUs x ~30 mins)	Thousands of samples, typically hundreds of files per sample	~1000	~35000	500 files/user
Access bandwidth required	500 Mbits/sec	1Gbit/s	0.5GB per job in less than 5 minutes	1 Gbit/s	Regular GEANT one is enough	~2GB per core (max)	10 Gbit/s	low	10 Gbit/s	low

Access pattern	Full files	All possibilities	Sequential read or write	Full files, Parallel access, R&W	Any, depends on the data (private/public (non) curated)	~1TB	Write once & read only for main data; R&W for products	Full files, R&W	Full files, R&W	R&W

Table 9: Storage requirements by thematic case.

Table 10 shows the consolidated information. Half of the cases require around 100GB of permanent storage with a granularity of a few thousand files. A bandwidth of 1Gb/s will suffice half of the cases.

	Min	Max	Total	Q1	Median	Q3
Permanent storage required	2 GB	500 GB	1 PB	20 GB	100 GB	200 GB
Granularity (number of individual files)	1	400,000	560.0	2,000	12,360	100,000
Access bandwidth required (MBytes/sec)	500 Mb/s	100 Gb/s	N/A	500 Mb/s	1 Gb/s	10 GB/s

Table 10: Summary of the storage resources requested (minimum, maximum, total and quartiles).

4. Validation of the Thematic services

4.1. Preliminary approach for the Validation

One of the objectives of EOSC-SYNERGY is to increase the relevance of the thematic services included in the project by adopting EOSC services to enable them to be exposed to a wider audience. Therefore, it is important to define how this increase will be measured.

For this purpose, we define in this deliverable the metrics to be used, the procedure to measure such metrics, the current values (baseline) and the expected values to achieve, whenever possible. Not all the services will evaluate all the metrics, as the way the service is operated may affect the relevance for such metrics. A summary table is provided by the end of the section.

4.2. Metrics to evaluate

We identify five categories for the metrics to be used:

- Impact on users.
- Impact on Capacity and Capability of the service.
- Impact on Scientific Outreach.
- Impact on the usability of the service.
- Impact on Cross-Fertilization.

Those metrics are defined in detail in the next subsections.

4.2.1. Metrics for the Impact on Users

The objective of these metrics is to evaluate the improvement on the service usage. The improvement could be relevant not only in the number of users, but also in the variety of them or their engagement. The metrics are shown in table 11.

Metric	Explanation	Objective	Units
MU_NUS	Number of different direct users who have accessed the service in a given period.	Maximize	Users/month
MU_NUSA	Accumulative Number of direct different users who have accessed the service since PM6.	Maximize	Users
MU_NIU	Number of different indirect[6] users who have accessed the service in a given period.	Maximize	Users/month
MU_NIUA	Accumulative Number of indirect different users who have accessed the service since PM6.	Maximize	Users

[6] An indirect user is a user that is not involved in the production of the data but consumes the results produced.

MU_NCEA	Accumulative Number of different centers where the users are based since PM6.	Maximize	Centres
MU_NCOA	Accumulative Number of different countries of origin of the users since PM6.	Maximize	Countries
NRUSA	Accumulative number of different users that accessed the service more than once since PM6.	Maximize	Users

Table 11: Metrics related to the number, activity and variety of the users.

The granularity of the user metrics along time will be on the order of "Users/month". As some of the thematic services may have a wider impact than the one of the individual users who run the simulations, a metric of indirect users could also be collected. Along with the number of users, it is important to observe the variety of such users in terms of centres and countries of origin.

4.2.2. Metrics for the Impact on Capacity and Capability

The objective of these metrics is to evaluate the improvement on the service capacity (HTC) and capability (HPC). The improvement could be relevant not only in the performance, but also in the different capabilities of the service. The metrics are listed in table 12.

Metric	**Explanation**	**Objective**	**Units**
MC_NSE	Number of service accesses in a given time frame.	Maximize	Accesses/month
MC_NSEA	Accumulated number of service accesses.	Maximize	Accesses
MC_CPU	Number of CPU hours used in a given time frame (per VCPU)	Maximize	CPU·hours[7]/month
MC_CPUA	Accumulated number of CPU hours used (per VCPU)	Maximize	CPU·hours
MC_MEM	RAM size used in a given time frame (hours used).	Maximize	GB·hours[8]/month
MC_MEMA	Accumulated number of CPU hours used (per VCPU)	Maximize	GB·hours/month
MC_STO	Number of CPU hours used in a given time frame (per VCPU)	Maximize	GB·hours
MC_STOA	Accumulated number of CPU hours used (per VCPU)	Maximize	GB·hours/month
MC_MXCC	Maximum capacity experimented (maximum number of VCPUs used simultaneously in production).	Maximize	VCPUS

[7] 1 CPU·hour is the usage of one VCPU in a cloud flavour during one hour. A VM with several VCPUs will multiply this CPU·hour cost accordingly. The granularity (e.g. at the level of the second, minute or hour) will depend on the monitoring system (the smaller the better).

[8] 1 GB·hour is the usage of one GB in a VM during one hour. The granularity (e.g. at the level of the second, minute or hour) will depend on the monitoring system (the smaller the better).

MC_MXCCP	Maximum computing capability experimented (maximum number of VCPUs used simultaneously for a single job)	Maximize	VCPUS
MC_MXMCP	Maximum memory capability experimented (maximum Memory size used simultaneously for a single job)	Maximize	GBs
MC_MXTHR	Maximum number of service accesses served simultaneously	Maximize	Accesses

Table 12: Metrics for evaluating the access to the services.

The metrics related to the accesses will evaluate the actual activity of the services and their alignment with the analysis of resource workload. Despite the objective being to maximize these numbers, it is important to avoid consuming unnecessary resources, so the ratios of some of the metrics could also be relevant (e.g. MC_NSE/MC_CPU). The metrics are also a good parameter for the evaluation of the infrastructure services in the project.

4.2.3. Metrics for the Impact on Scientific Outreach

The objective of the thematic services of EOSC-SYNERGY is to serve the scientific community with data and processing services to advance in the research activities. Despite that this will be difficult for the timeline of the project, this set of metrics try to address the scientific interest of the thematic service. The metrics are listed in table 13.

Metric	**Explanation**	**Objective**	**Units**
MO_PUB	Number of publications acknowledging the service.	Maximize	Publications
MO_COM	Number of communications (talks, panels, posters, etc.) acknowledging the service.	Maximize	Communications
MO_TRAH	Number of individual training hours on the service. For each training action, compute the hours (or fractions) of the sessions with participation of the service and multiply by the number of trainees.	Maximize	trainee·hours

Table 13: Metrics for evaluating the outreach of the services.

The evaluation of the scientific outreach is a long-term metric that can also help to recognize the relevance of individual reports. However, some user communities may not be used to cite or acknowledge services used during their research activities.

4.2.4. Metrics for the Impact on Usability

The usability of the thematic services can be assessed at different dimensions, following an usability questionnaire. For this purpose, we will define a pair-based evaluation involving external and internal users as much as possible. The information will be collected through the standardised questionnaire shown in table 14.

Metric	Feature	Verification[9]	Rank[10]	Comment
MU_PER	Performance			
MU_ERR	Error management			
MU_SCA	Scalability			
MU_COM	Completion			
MU_INT	Interoperability			
MU_LC	Learning curve			
MU_CON	Convenience			
MU_ROB	Robustness			
MU_OVA	Overall assessment			

Table 14: Metrics for the evaluation of the usability of the thematic services.

The evaluation of the usability is a good metric for evaluating the quality of the services from the technical point of view. It provides relevant information for the application developers who could find the weaknesses of the services from the technical point of view.

4.2.5. Metrics for the Impact on Cross-fertilization

The interest of European projects is to maximize the collaboration among centres from different entities and countries. Therefore, it will be important to measure the transfer of knowledge among the thematic services. The metrics are listed in table 15.

Metric	Explanation	Objective	Units
MF_COSH	Number of code transfers, measured on the adoption of base containers, github forks, source code templates, etc.	Maximize	Transfers
MF_JDIS	Number of joint dissemination actions (publications, communications or joint sessions on events).	Maximize	Actions
MF_SYN	Number of synergies among thematic services, as the sum of COSH, JDIS and any other action not reflected in any of these metrics.	Maximize	Synergies

Table 15: Metrics to evaluate the cross-fertilization actions among the thematic services.

[9] How this feature has been verified (e.g. through experimentations, through trustworthiness metrics, through analysis...)

[10] Rank: 1 Missed; 2 Achieved below expectancies; 3 Achieved as expected; 4 Achieved above expectancies

4.3. Metrics Gathering procedure

The metrics stated in the previous section will be gathered directly from the thematic service or indirectly through other monitoring and accounting services.

In the case of the user metrics, this will imply the need to extract specific information from the authentication and authorization services and monitoring. Table 16 summarises how such metrics can be obtained for each one of the thematic services.

	WorSiCa	G-Core	SAPS	OpenEBench	Scipion	LAGO	SDS-WAS	UMSA	MSWSS	O3AS
MU_NUS	VO+group / WorSiCa service	EO mission owners and Web application front-end	VO+group / SAPS service	VO+group / OpenEBench service	N/A ScipionCloud service	VO+group / LAGO service	SDS-WAS service	VO + group	MSWSS service	N/A
MU_NUSA	VO+group / WorSiCa service	Same as declared in MU_NUS	VO+group / SAPS service	VO+group / OpenEBench service	N/A ScipionCloud service	VO+group / LAGO service	SDS-WAS service	VO + group	MSWSS service	N/A
MU_NIU	N/A	>100, contact EO missions	TBD	TBD	TBD	Inquiring users + Anonym. access	TBD	inquiring users	inquiring users	N/A
MU_NIUA	N/A	Unlimited	TBD	TBD	TBD	Inquiring users + Anonym. access	TBD	Inquiring users	inquiring users	N/A
MU_NCEA	VO+group / WorSiCa service		VO+group / SAPS service	VO+group / OpenEBench service	N/A EGI / ARIA Login?	VO+group / LAGO service	SDS-WAS service	by user affiliation	by user affiliation	VO+group / LSDF service
MU_NCOA	VO+group / WorSiCa service	world wide open	VO+group / SAPS service	OpenEBench stats + ELIXIR AAI for registered users	Scipion statistics EGI / ARIA Login?	VO+group / LAGO service	SDS-WAS service	by user affiliation	by user affiliation	VO+group / LSDF service
NRUSA	VO+group / WorSiCa service	Undefined	SAPS service	TBD	N/A	VO+group / LAGO service	SDS-WAS service	possible from accounting logs	MSWSS service	N/A

Table 16: Procedure for obtaining the measures for the metrics.

Table 17 shows the procedures for obtaining the metrics related to the accesses to the service and the usage of resources. Clearly, most of the metrics related to the use of resources will require obtaining information from the accounting services. The second source for obtaining the rest of metrics will come from the access portals.

	WorSiCa	G-Core	SAPS	OpenEBench	Scipion	LAGO	SDS-WAS	UMSA	MSWSS	O3AS
MC_NSE	WorSiCa service	> 5	SAPS service	OpenEBench service	ScipionCloud service	LAGO service	SDS-WAS service	access logs of web FE.	MSWSS service	LSDF service
MC_NSEA	WorSiCa service	N/A	SAPS service	OpenEBench service	ScipionCloud service	LAGO service	SDS-WAS service	access logs of web FE.	MSWSS service	LSDF service
MC_CPU	Accounting	N/A	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting
MC_CPUA	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting
MC_MEM	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting
MC_MEMA	Accounting	N/A	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting
MC_STO	Accounting	N/A	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting
MC_STOA	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting
MC_MXCC	N/A	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting
MC_MXCCP	N/A	Accounting	Accounting	Accounting	Accounting	N/A	Accounting	Accounting	Workload manager	Accounting
MC_MXMCP	N/A	N/A	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	Workload manager	Accounting
MC_MXTHR	N/A	Accounting	Accounting	Accounting	Accounting	Accounting	Accounting	access logs of web FE.	MSWSS service	Accounting

Table 17: Sources for obtaining the metrics related to access and resource usage.

The metrics related to the usability will be collected by interviewing key users and by means of the questionnaire shown in table 14.

The metrics related to the Scientific Outreach will be obtained by requesting the information to the users and the training activities, as reflected in table 18. Finally, the metrics related to cross-fertilization will be obtained globally inside WP4.

	WorSiCa	G-Core	SAPS	OpenEBench	Scipion	LAGO	SDS-WAS	UMSA	MSWSS	O3AS
MO_PUB	inquiring users	Inquiring EO missions	inquiring users	inquiring users	Inquiring users / ARIA	LAGO repository	inquiring users	users' yearly reports	inquiring users	inquiring users
MO_COM	inquiring users	Inquiring EO missions	inquiring users	Inquiring users	Inquiring users	LAGO repository	inquiring users	users' yearly reports	inquiring users	inquiring users
MO_TRAH	WorSiCa team	G-Core team	Through WP6	OpenEBench team	Scipion team	Through WP6	SDS-WAS team	Through WP6	MSWSS team	Through WP2 and WP6

Table 18: Outreach metrics collection procedures.

5. Data Management Plans

Due to the nature of EOSC-SYNERGY, Data Management Plans should be considered individually at the level of each thematic service. As the adaptation of the thematic services will surely imply changes on the DMPs, we consider that all DMPs will be living documents that will be amended, improved and detailed along the project timeline. Therefore, DMPs should have a clear version number and include a timetable for updates.

The DMPs have been defined according to the template of the DMP Online tool[11]. This template proposes six sections:

1. Data summary, addressing the purpose of the data collection, the types of formats, the origin of the data, its size and how the data will be reused.
2. The FAIR-ness of the data produced, addressing:
    a. Findability: Data discoverability, identifiability, naming conventions, use of DOIs, versioning and metadata standards.
    b. Accessibility: Which data will be open and how, software tools for accessing, associated metadata and documentation, access restrictions.
    c. Interoperability: Standards and vocabularies for data and metadata used.
    d. Reusability: Licensing, reusability conditions, quality assurance and data validity.
3. Allocation of resources, with the estimation of the cost for making the data FAIR, responsibilities for data management and long term preservation.
4. Data security, in the case of managing sensitive data.
5. Ethical aspects, also in the case of data with ethical implications.
6. Any other regulations you should consider.

A summary of the DMPs from all the thematic services is provided in the next sections.

5.1. Data Summary

The thematic services of EOSC-SYNERGY are different in objective, area and structure. Therefore, the definition of the Data Management Plans (DMPs) should be done individually. The summary section of the DMPs provide a good insight of the data sources, targets and objectives of the data collection. Table 19 shows a consolidated view of the summary section of the ten DMPs focusing on the data formats, whether the data is reused from other services, the origin of such data, its size and the targets.

Most of the thematic services consume data from public or community repositories and all of them encode data produced on a standardized format. In some cases the thematic services consume user-specific data with restricted access. The data size has been already considered in the analysis of the resources required and the target sources cover different scientific and engineering disciplines.

[11] https://dmponline.dcc.ac.uk/

	Data Formats	Data Reusability	Data Origin	Data Size	Target
WORSICA	Open Geospatial Consortium (OGC)	Yes	ESA/Copernicus, EMODnet, EOSC-hub OPENCoastS service, Pleiades satellite data provider, GEBCO bathymetry, Smith & Sandwell Topography and also in situ and UAVs data	20 TB	Researchers, civil protection authorities, environmental agencies and water management authorities
G-Core	ISPs, zip packages, tiff, JPEG2000	Yes	Satellite data (e.g. Coopernicus, SMOS, PAZ)	Tens of GB	Wide range of researchers
SAPS	INSPIRE, OGC SOS, EUOSME	Yes	LANDSAT, NCEI, CSI	108 TB	Researchers in Agriculture Engineering and Environment.
OpenEBench	Standard Bioinformatics data formats, such as FASTA, FASTQ, BAM, VCF, etc.	Yes	Different sources, current challenges come from DREAM (Dialogue on Reverse Engineering Assessment and Methods), the Cancer Genome Atlas (TCGA) and the Quest for Orthologs (QfO) challenge.	Up to MBs	Life Sciences scientific communities
Scipion	Common EM image formats (mrc, tiff, hdf, em, spi, vol, map and others)	Yes	Cryo-Electron Microscopes	up to TBs	Structural Biologists
LAGO	CORSIKA outputs, other formats may be considered	Yes	Latin American Giant Observatory (LAGO)	>156TB	Astrophysics community mainly but also High Energy Physics, Life Sciences, Weather Forecasting, Aerospatial security or Computer Science
SDS-WAS	netCDF	Yes	SDS-WAS partners (BSC, NASA, NCEP, ECMWF, etc ...)	>4TB	Researchers, meteorological agencies, enterprises (solar, aviation, ...)
UMSA	mzML and JSON for metadata	No	Mass spectrometers in the laboratories of the users	Hundreds of TBs	EIRENE ESFRI target users
MSWSS	GIS, CIS and SCADA standard formats	Yes	Own data plus other sources (ZBGIS, OpenStreetMap, DEM50)	~20GB	Researchers on Water Networks distribution
O3AS	NETCDF	No	IGAC/SPARC Chemistry Climate Model Initiative provided by the CEDA and ERA-Interim or ERA5	200TB	Scientists working on the Ozone Assessment report

Table 19: Data Management Plans, Summary section.

5.2. Data FAIRness

The second relevant section analysed is the FAIR (Findability, Accessibility, Interoperability and Reusability) of the data produced by services. Table 20 shows a few key aspects of these FAIR properties. In Findability, we analyse the metadata formats used and the usage of DOIs. In the Accessibility, we identify if the service has already identified a license model for the data produced and how the data will be accessed. In the Interoperability, the table shows if there are ontologies used in the representation of the data, and in the Reusability we present if there are data quality techniques applied to the data and the embargo period for releasing publicly the data.

	Findability		Accessibility		Interoperability	Reusability	
	Metadata Formats	DOIs	License	Methods	Ontologies	Data Quality	Embargo
WORSICA	INSPIRE, OGC SOS, EUOSME	Dataverse or ZENODO	Freely available	WORSICA portal	Metadata, keywords, SI units	Process chain documented	Yes
G-Core	OGC CSW-ebRIM	No	Depending on the source	Embedded in the GCORE services and Marketplace	ISO 19115	Not explicitly	Depending on data
SAPS	PROV	No	Open to all registered users	SAPS portal	No	Not automatic	No
OpenEbench	JSON	Through EUDAT or ZENODO	Different	Through the OpenEBench Portal	EDAM	Internal processing	TBD
Scipion	JSON (CWLprov)	EMPIAR/EMDB (EBI Databases)	Open after embargo (3 years) (CC By)	Through the EMPIAR/EMDB database portals	N/A	Workflow documented	Yes
LAGO	Through B2FIND	Through B2Handle	Open after embargo (BSD-3 or CC)	B2FIND and CORSIKA tools	Dublin Core, CORSIKA specifications	Managed by Providers	Yes
SDS-WAS	netCDF (CF-1.6)	TBD	Open to all registered users (if the agreement is signed)	SDS-WAS portal	NetCDF and CF convention	Managed by Providers, workflow documented	No
UMSA	JSON	Own DOIs	TBD	Through the UMSA service	TBD	TBD	TBD
MSWSS	JSON	Own DOIs	Depending on data	Through the MSWSS service	TBD	Not automatically	Depending on data
O3AS	CF (Climate and Forecast)	KIT or HDF will provide DOI	Apache-v2, an MIT or a GPL license	O3AS interface and via git, netCDF data viewers	NETCDF and CF convention	Assured by the data providers	No

Table 20: Data Management Plans, FAIR evaluation section.

Most of the thematic services have already identified the way they will annotate the data to be findable, as well as the license model used. The access to the data will be mainly performed directly through the thematic service. No data quality procedures are applied and most of the thematic services already foresee an embargo period for the data.

5.3. Other aspects

In this last section, we have consolidated in table 21 some information related to the sections 3 to 6 of the DMP template. Basically, we analyse if the thematic service already has computed the costs for ensuring the FAIRness of the data or if they have identified the sources for the resources required, as well as the ethical requirements and security means applied to the data. Finally, we compile the limitations that will apply to the FAIRness of the data.

	Resource Allocation	Ethical Requirements	Security	Limitations
WORSICA	Partially local and external	Article 34 of the Grant Agreement	Dataverse or ZENODO	1 month preservation of user's data
G-Core	Marketplace co-funding, support from missions	N/A	Registered users	N/A
SAPS	UPV + UFCG, exploring sustainability	N/A	Registered users	N/A
OpenEbench	ELIXIR ERIC	Yes, clearly managed	ELIXIR AAI	N/A
Scipion	Instruct ERIC will evaluate long term sustainability.	N/A	Registered users	N/A
LAGO	Over 50K€	N/A	Registered users	N/A
SDS-WAS	Internal	N/A	Registered users	N/A
UMSA	TBD	TBD	TBD	N/A
MSWSS	TBD	N/A	Registered Users	Access to operational data can be limited by national legislation or institutional policies
O3AS	KIT and HDF plus funding of the Helmholtz Programme	N/A	Freely accessible	Limited by national legislation / institutional policies

Table 21: Data Management Plans, other aspects section.

In this final analysis, most of the services have not yet identified the cost or the sources of funding for ensuring the FAIRness of the data. Ethical implications are not applicable in most of the cases, as the thematic services do not deal with sensitive or personally identified data. Data security means applied to the data access is typically implemented through the thematic services or the long-term repositories. Finally, some of the services already have identified limitations on the time that the users' data will be kept as well as some regulations that apply to the produced data.

6. Conclusions

This document presents a detailed analysis of the ten thematic services of EOSC-SYNERGY. The ten thematic services present several differences and commonalities, which increases their relevance and complementarity.

The document has first described the thematic services identified their gaps and bottlenecks, as well as the plans for the adaptation of the thematic services to be integrated in the EOSC ecosystem. The thematic analysis reveals a few technical solutions that are relevant to most of the thematic services. These services will be analysed to define best practices that will be of interest for all the thematic services.

A second important aspect of the deliverable is the analysis of resources expected to be needed during the project lifetime. Some of the thematic services are fairly intense on computing and most of them are intense in data storage. This characterization is key to evaluate the rightmost services to be used also.

A third aspect covered by the deliverable is the definition of the metrics for the Key Performance Indicators that will be used along the project to evaluate their performance. Five groups of metrics, related to users, service access, usability, scientific outreach and cross fertilization have been defined.

Finally, the deliverable goes through the Data Management Plans for the Thematic Services. As those ten thematic services cover different areas, are managed by different groups and involve different users, the DMPs must be defined individually at the level of each Thematic Service.

The deliverable includes detailed annexes for the resource and technical analysis and the DMPs for each one of the thematic services.

A. Annex - Detailed Technical Analysis

The annexes of the deliverable include more detailed information about the technical services and the workload requirements collected during the writing of the deliverable. This information is included in the deliverable for further reference.

A.1. WorSiCa

A.1.1. Technical solution

This section describes the basic technical services needed to build the thematic service.

http://www.dha.lnec.pt/worsica/				
Service Scope	**Service used**	**Limitation**	**EOSC service?**	**Provider**
AAI	Now: Local Plan:EGI Check in		Yes	EGI Federation
Workload Mng.	Now: Local batch system Plan: HT Condor or EGI Workload Mng.	Need Accounting	Possibly	
Resource Mng.	Now: Manual Plan: IM, Ansible		Yes	EOSC-hub marketplace
Data Storage	Now: Local Plan: Object storage or EGI Data-hub		Yes	EOSC-hub marketplace
Monitoring	Now: N/A Plan: ARGO	Need to develop probe	Yes	EOSC-hub marketplace
Other: Hydrodynamic water forecasts	OPENCoastS		Yes	EOSC-hub marketplace
Computing Resources	Now: Local Plan: FedCloud and EGI HTC		Yes	Now: INCD Plan: IberGrid
Storage Resources	Now: Local Plan: FedCloud and EGI Online storage		Yes	Now: INCD Plan: IberGrid

Table 22: Technical solutions used and on plan for WorSiCa.

A.1.2. Data and Workload Analysis

The WorSiCa service generates a set of several products that make use of different application models, such as batch jobs (Parallel (including GPUs) and High Throughput) and Interactive applications. Applications could be both CPU and Data-Intensive.

The service can be operated at three typical sizes: small, medium and large:

- For **small** deployments, we use as case the coastal processing with one satellite image set;

- For **large** deployments, we use as case the water leak processing using all available satellite image sets from 08-2016 till today and from different orbits (assume 402 image sets and still increasing).
- Average of both cases is the **medium** estimated size value.

Resources expected to be requested by WorSiCa to achieve the metric thresholds.

Computing	Total CPU Consumption by time	**For one user request** **Small:** 1 hours/week (WorSiCa) + dependent services (OPENCoastS from EOSC-Hub) **Medium:** 200 hours/week (WorSiCa) + dependent services (OPENCoastS from EOSC-Hub) **Large:** 400 hours (WorSiCa) + dependent services (OPENCoastS from EOSC-Hub) **For the posterior uses on requests** 3 hours/week (WorSiCa) + dependent services (OPENCoastS from EOSC-Hub)
	Number of individual jobs	Two jobs per user per week
	RAM required by each single job.	**Small:** 4 GB (WorSiCa) + dependent services (OPENCoastS from EOSC-Hub) **Medium**: 8 GB (WorSiCa) + dependent services (OPENCoastS from EOSC-Hub) **Large:** 16 GB (WorSiCa) + dependent services (OPENCoastS from EOSC-Hub)
	Storage required by each single job.	**Small:** ~1.6GB (WorSiCa) + dependent services (OPENCoastS from EOSC-Hub) **Medium:** 256GB (WorSiCa) + dependent services (OPENCoastS from EOSC-Hub) **Large:** ~ 550GB (WorSiCa) + dependent services (OPENCoastS from EOSC-Hub)
	How does it run?	Variable demand: 40-50 runs/month
	Size of the input data per job	**Small:** ~1,1GB (1 imgset) **Large:** ~550GB
	Size of the Output data per job:	**Small:** ~500MB **Large:** ~10GB
Storage	Permanent storage	20 TB
	Granularity (number of individual files)	**For each job** **Small:** 10, **Large:** ~2000 individual files
	Access bandwidth	500 MBits/sec
	Access pattern	Full files
Network	Does it need to be accessed from outside?	Yes
	Does it require data from external sources?	- Volume to transfer: Small: ~1,1GB (1 imgset); Large: ~550GB; - Bandwidth: 500 MBits/sec

Table 23: Resource profile of WorSiCa

A.2. G-CORE

A.2.1. Technical solution

This section describes the basic technical services needed to build the thematic service.

Service Scope	Service used	Limitation	EOSC service?
AAI	user/pwd with SSO and Kerberos LDAP and CAS	Inherent to each technology applied and dependent of type of mission	Possibly
Workload Mng.	GCore+ K8s	Inherent to each technology applied and commercial agreement to reach in this context.	Possibly
Resource Mng.	Rancher + IM/EC3	Inherent to each technology applied	NA
Data Storage	ElasticSearch for the catalogue	Inherent to each technology applied	NA
Monitoring	GCore	Inherent to each technology applied and commercial agreement to reach in this context.	Possibly
Computing Resources	Now on-premise (depending of mission) Next on-cloud	Inherent to each technology applied and mission constraints	
Storage Resources	Now on premise (NFS, SAS) Next S3 Object Storage	Inherent to each technology applied and mission constraints	

Table 24: Main Technical services of G-core.

A.2.2. Data and Workload Analysis

Depending on the type of mission, different performances are required. In general, the missions use a sequential processing chain with a high throughput. Currently there are no parallel jobs for the same products but it is envisaged to include data cube processing for some type of sensors that fits this possibility. The main limitations are CPU-intensive use for processing tasks (the systems usually are dimensioned to reach 80-90% of CPU load), large capacity of storage for the entire mission archive and fast access to them. In addition, it is important to move the data from the system to the users in a fast way. Another important feature is that these systems are very intensive in occupation for processing the activities usually last for many hours per day. The application is in Production and under upgrade.

Resources expected to be requested by WorSiCa to achieve the metric thresholds.

Computing	Total CPU Consumption by time	32 cores with an occupation of 60% of time per day
	Number of individual jobs by time	100 jobs/day for processing
	RAM required by each single job.	2GB to 16GB[12]
	Storage required by each single job.	0.001GB to 8GB
	How does the application run?	Periodically every satellite passes. Typically four times per day.
	Size of the input data per job	0.1MB to 2000MB
	Size of the Output data per job:	0.1MB to 8000MB
Storage	Permanent storage required (GBytes)	The entire mission. E.g. PAZ 36 TB in disk the rest goes to tape. Sentinel LTA is in XXPB
	Granularity (number of individual files)	1000-100000
	Access bandwidth required (MB/sec)	1Gbit/s
	Access pattern	All possibilities (Full files, partial files, single access, repeated access, Read only, R&W)
Network	Does the application need to be accessed from outside?	No
	Does the application require data from external sources?	- Products size around 1MB to 8 GB. - Auxiliary data for processing around 100sMB

Table 25: Resource profile of G-Core

[12] A range interval is given depending of the mission

A.3. SAPS

A.3.1. Technical solution

This section describes the basic technical services needed to build the thematic service.

SAPS Endpoint	**http://demo.saps.lsd.ufcg.edu.br**		
Service Scope	**Service used**	**Limitation**	**EOSC service?**
AAI	Local authorisation token	Own system, limited to the application, duplication of credentials	No
Workload Mng.	Own Scheduler	Limited to the SEBAL pipeline.	No
Resource Mng.	Fogbow	Not automated scalability.	No
Data Storage	OpenStack Swift		No
Monitoring	None		No
Other			
Computing Resources	OpenStack Cloud @ UFCG		Compatible with EOSC EGI Cloud Compute
Storage Resources	OpenStack Cloud @ UFCG		No

Table 26: Main Technical services of SAPS.

A.3.2. Data and Workload Analysis

The application type is mainly Sequential or High-Throughput, limited by CPU and RAM. The application is mature and exposed as an advanced prototype among experts in the area.

Computing	Total CPU Consumption by time	200KCPU/months (per experiment)
	Number of individual jobs by time	4.000 jobs/month (per experiment)
	RAM required by each single job.	16G bytes
	Storage required by each single job.	~4G bytes
	How does the application run?	Each job computes evapotranspiration and other vegetation indexes of an entire Landsat scene (170km x 185km, 30-meter resolution) for a particular date. A job

		represents a three-step sequential workflow: input download, input preprocessing, and evapotranspiration estimation. Users submit batches of jobs typically for a single scene and a period of 40 years, which leads to a batch of about 800 jobs.
	Size of the input data per job	0.4 GB
	Size of the Output data per job:	~4 GB
Storage	Permanent storage required	4GB per job
	Granularity (number of individual files)	100 file per job
	Access bandwidth required	0.5GB per job in less than 5 minutes
	Access pattern	Input full files (download phase), output generated at the end of each step of the pipeline (sequential read or write, depending on whether it is an input or output file).
Network	Does the application need to be accessed from outside?	Yes.
	Does the application require downloading data from external sources?	Yes. Around 0.5GB per job [downloading should be done in less than 5 minutes].

Table 27: Resource profile of SAPS.

A.4. OpenEBench

A.4.1. Technical solution

Endpoint:	**https://openebench.bsc.es**		
Service Scope	**Service used**	**Limitation**	**EOSC service?**
AAI	Current: ELIXIR AAI Planned: Life Sciences AAI		Yes
Workload Mng.	Local: Make use of SGE Planned: Make use of the GA4GH TRS/WES/TES stack.	SGE will be decommissioned. We cannot submit jobs elsewhere with the current configuration	No
Resource Mng.	OpenNebula		No
Data Storage	Current: Local Planned: We will make use of Zenodo/EuDat to manage data sets.		Zenodo/EuDat are EOSC services.
Monitoring	UpTime robot	External solution for monitoring the whole infrastructure availability	No
Computing Resources	Local: Current instance runs at StarLife (BSC). Planned: Making use of resources available at EOSC-Life	Computation time allocation might be an issue depending on the communities requirements.	Yes
Storage Resources	Local. Planned: Data sets will be exported to Zenodo/EUDAT and/or other repositories defined by the communities. Workflows will be deposited in WorkflowsHub (from EOSC-Life)		Yes

Table 28: Technical services identified for OpenEBench.

A.4.2. Data and Workload Analysis

Applications running on the service can be of different types, including shared memory multithreaded, distributed Memory, High Throughput and sequential. The application is in production and it can be intense in CPU, Memory or I/O, depending on the actual benchmark run.

Computing	Total CPU Consumption by time	Depending on the benchmarking workflow. Current usage of level 2 is about 0.1KCPU hours/week
	Number of individual jobs by time	10 jobs/week
	RAM required by each single job.	Depending on the benchmarking workflow. Current communities, from 2GB to 8GB
	Storage required by each single job.	Average of 150MB (in logs files, temporary data and results)
	How does the application run?	Web-based petitions. It runs periodically depending on communities' needs. In the case of concurrent challenges we can expect up to 100 weekly executions.
	Size of the input data per job	Depending on the benchmarking workflow. Current communities, ~70MB
	Size of the Output data per job:	Average of 100Mb
Storage	Permanent storage required	100 Gb. Data of unregistered users is periodically cleaned.
	Granularity (number of individual files)	Average of 500 files/user
	Access bandwidth required	Not Available.
	Access pattern	R&W
Network	Does the application need to be accessed from outside?	Yes
	Does the application require downloading data from external sources? Yes	Yes. We do mirror locally most used files from external repositories. Once those files are not anymore used, we keep a reference and remove them from our local filesystem. This is very much dependent on scientific communities but usual cases range from 1 to 10 Gb. We are discussing the possibilities to incorporate the metagenomics community that in total uses roughly 1PB of storage for their activities.

Table 29: Resource profile of OpenEBench.

A.5. Scipion Cryo-Electron Microscopy Service

A.5.1. Technical solution

Current implementation, not yet in production, is based on a set of services but the plan is to upgrade it to use different services, probably EOSC services or INDIGO-DataCloud recommendations.

Scipion endpoint:	**Not yet in production**		
Service Scope	**Service used**	**Limitation**	**EOSC service?**
AAI	Now: Westlife AAI (AARC based). Plan: EGI Checkin.	Instruct users have ARIA accounts (EGI AAI supports it). Need to access EGI compute.	Plan Yes.
Workload Mng.	Now: torque. Plan: CLUES + slurm / K8s (docker needed).	Better optimization of resources based on steps HW requirements.	
Resource Mng.	Now: Cloudify. Plan: IM.	Cloudify occi plugin (occi no longer supported).	
Data Storage	Local + OneData.	See section below.	Yes.
Monitoring	None.	Not real usage info available.	Is there any?
Other	Now: Remote desktop: TurboVNC+VirtualGL+noVNC	Require remote GPU.	
Computing Resources	Now: EGI FedCloud. Plan: EOSC compute cloud and AWS EC2.		
Storage Resources	Block storage and OneData.	OneData and NFS. I/O performance.	Yes.

Table 30: Technical services required by Scipion.

A.5.2. Data and Workload Analysis

The application has an interactive GUI to manage projects and workflows that run batch shared-memory multithreaded and GPU-enabled applications. Some of the workflow steps require user intervention through the GUI.

The application is limited by many factors such as CPU, Data, GPU availability and heterogeneous steps. It has been a mature application in production for many years. The Scipion on demand in the cloud is under development.

CryoEM workflows are highly heterogeneous in terms of computational resources, where workflow steps can be seen as the minimum unit for processing. Numbers below correspond to a standard workflow (jobs = workflow steps although each step launches a high number of batch jobs depending on parallelization and input data). This processing can take weeks.

Computing	Total CPU Consumption by time	Highly variable but a minimum of 0.4KCPU hours/week.
	Number of individual jobs by time	Around 100 workflow steps, some of them sending hundreds of jobs.
	RAM required by each single job.	Depending on the step, some require up to 32 GB.
	Storage required by each single job.	If movies are not considered the most demanding step would require around 100 GBs. Others much less (~ 2 GB average).
	How does the application run?	Server deployment for exclusive use for 2 weeks. Server remote access through the web browser.
	Size of the input data per job	If movies are not considered the most demanding steps would require around 100 GBs of input data. Others much less (~ MBs or few GBs).
	Size of the Output data per job:	If movies are not considered the most demanding steps would produce around 100 GBs of output data. Others much less (~ MBs or few GBs).
Storage	Permanent storage required	If movies are not considered the whole project might be around 1 TBs max.
	Granularity (number of individual files)	Thousands.
	Access bandwidth required (MBytes/sec)	No special requirements but data transfer slower.
	Access pattern	Full files, parallel access, R&W.
Network	Does the application need to be accessed from outside?	Remote browser for user interaction through GUI (noVNC +VirtualGL on GPU powered machine for 3D rendering)
	Does the application require downloading data from external sources?	No but data has to be preloaded before application can run. However, If movies are not considered then micrographs and the preprocessing project should be transferred beforehand (order of GB's). For OneData disk sharing connectivity is also important to guarantee processing.

Table 31: Resource profile of Scipion.

A.6. Latin American Giant Observatory - LAGO

8.6.1. Technical solution

LAGO Endpoint:	Collaboration: http://lagoproject.net/ (it will link the thematic service). Specific page for the thematic service: not yet.		
Service Scope	**Service used**	**Limitation**	**EOSC service ?**
AAI	Now: Username and password Plan: EGI Checkin.	(see data storage lim.)	No, to be integrated within the project.
Workload Mng.	Now: Local cluster batch Plan: Cluster batch & (EC3 or EGI Workload S.)	(see data storage lim.)	
Resource Mng.	Now: Local cluster batch Plan: Cluster batch & IM+EC3/VMops	(see data storage lim.)	
Data Storage	Now: Local filesystems Plan: B2 data tools (B2FIND, B2HANDLE) & EGI DataHub (i.e OneData)	- B2 & DataHub (mainly) should work fine on private clusters (compatibility). - Data confidentiality before waiting period.	
Monitoring	Now:N/A Plan: ARGO	Some accounting and monitoring is needed for continuous processing raw to analysis	
Computing Resources	Now: ACME cluster at CIEMAT Plan: Local Cluster &	To support the continuous processing: ~ 40- 80 cores/day available.	
Storage Resources	Now: Storage servers at CIEMAT Plan:B2 data tools (B2FIND, B2HANDLE) & EGI DataHub (i.e OneData)	Estimated: 5.6-11.6 TB/year	

Table 32: Technical services identified for LAGO Thematic Service.

8.6.2. Data and Workload Analysis

The analysis of the data and workload requirements will be focused on CORSIKA, the main application of LAGO. This application is mostly High Throughput (sequential with many individual and independent jobs) and mainly limited by CPU. It is a mature and in-production application.

Computing	Total CPU Consumption by time	(As orientation, a single *3,600s background simulation* longs for 17 h on 40 cores Intel Xeon Gold 6138 CPU @ 2.00GHz) 1 pre-processing or analysis job = 0.6 CPU hours Real data (1 month, 1 WCD) = 302.4 CPU hours/month = 70.56 hours/week 1 simulation job = 0.25 - 10 CPU hours

		1 Background *S*imulation (*3,600s*) = 60 simulation jobs = 680 CPU hours **Total** 4 WCDs + 25 users (1 sim./month) = **4,248.9 CPU hours/week (~ 26 cores/day)**
	Number of individual jobs by time	Pre-processing and analysis real data (1 month, 1 WCD)= 504 jobs/week Simulation (25 users, 1 sim./month) = 350 jobs/week **Total** 4 WCDs + 25 users (1 sim./month) = **2,366 jobs/week (338 jobs/day)**
	RAM required by each single job.	4GB (maximum)
	Storage required by each single job.	Temporal space in VM, container or node (input and output decompressed): - Pre-processing: ~ 10 GB - Analysis: ~ 6.5 GB - 1 *background* simulation job (*60s*): 0.1-10GB - 1 *event* simulation (*60s*): 1-100GB **Usually: 10GB**, rarely 100GB (event simulations)
	How does it run?	- Real data: single large experiment associated with each detector. - Simulations: variable demand, ~ 1,500 jobs/month (25 users, 1 sim. /month)
	Size of the input data per job	- Pre-processing: ~ 200 MB (one file compressed) - analysis: ~ 100 MB (one file compressed) - Simulations: order of KB (one file compressed)
	Size of the Output data per job:	- Pre-processing: ~ 100 MB (one file compressed) - Analysis: ~ 30 MB (one file compressed) - *Background* simulation 2MB-200MB (two files compressed) - *Event* simulation: 20MB-2GB (two files compressed) **Usually: 100-200MB,** rarely 2GB.
Storage	Permanent storage required	Minimum (4 WCDs + 25 users): 5.6 TB/year Maximum(4 WCDs + 25 users): 11.6 TB/year
	Granularity (number of individual files)	**Total** 4 WCDs + 25 users(1 sim./month): **157,680 files/year (in 103,980 reachable data-sets)** **Size**: 50KB - 2GB files (**usually 100-200MB** compressed files)
	Access bandwidth	Regular GEANT one is enough.
	Access pattern	All, depending on the data (private non curated, private curated, public)
Network	Does the application need to be accessed from outside?	Not to the moment, but it is a possibility

	Does the application require downloading data from external sources?	Yes, LAGO repositories

Table 33: Resource profile of LAGO Thematic Service.

A.7. Sand and Dust Storms Warning Advisory and Assessment System - SDS-WAS

A.7.1. Technical solution

Service Scope	Service used	Limitation	EOSC service?
AAI	Web login with username and password		
Workload Mng.	Local		
Resource Mng.	Local		
Data Storage	Local		
Monitoring	Local nagios		
Other			
Computing Resources	BSC clusters	Queues sometimes	
Storage Resources	GPFS + not parallel archive (long term data)		

Table 34: Technical services identified for SDS-WAS Thematic Service.

A.7.2. Data and Workload Analysis

The model job part of the application currently works as an HPC distributed-memory parallel application. The application will also include an interactive web service to download data from a web GUI or an API. Data visualization and data analysis are partially interactive. Users can browse through predefined images and numerical scores pre-generated/calculated offline. The application is currently in production (except for the interactive part). The application is bounded both by CPU and Data requirements.

Computing	Total CPU Consumption by time	878 CPU hours/week
	Number of individual jobs by time	1 daily job (248 CPUs x ~30 mins)
	RAM required by each single job.	~2GB per core (max)

	Storage required by each single job.	~1TB
	How does the application run?	1 daily execution
	Size of the input data per job	~50 MB
	Size of the Output data per job:	~900 MB
Storage	Permanent storage required	Currently is 1GB per day, but more variables will be added
	Granularity (number of individual files)	1 daily file
	Access bandwidth required	A wished feature could be to have bandwidth control per user/session to avoid stuck the service
	Access pattern	Full files, aggregated files, Read only
Network	Does the application need to be accessed from outside?	Yes
	Does the application require downloading data from external sources?	Yes

Table 35: Resource profile of SDS-WAS.

A.8. UMSA: Untargeted Mass-spectrometry Analysis

A.8.1. Technical solution

Long-term storage of the experimental data is implemented as a conventional, filesystem based system, internally running GPFS over a cluster of storage servers, and exposing NFS, CIFS, and FTP interfaces. The service is operated by Masaryk University, in collaboration with CESNET.

Currently the same solution is applied for secondary data (results of computational analyses), but we expect to migrate to object storage due to increasing data sizes (reaching PB scale). Ceph storage cluster operated by CESNET will be used.

Workflow management is implemented by Galaxy (http://getgalaxy.org) software running in a virtual cluster at the CESNET/MU Openstack site (http://cloud2.metacenturm.cz). Workload inside the cluster is managed by a dedicated batch system (Slurm).

Key application software components are apLCMS (https://sourceforge.net/projects/aplcms/), xMSAnalyzer (https://sourceforge.net/projects/xmsanalyzer/) and erah (https://cran.r-project.org/web/packages/erah/index.html). These software modules are wrapped as tools in Galaxy.

In parallel, various other software pipelines (MS-DIAL, mzMine, MetAlign, MSCS, ...) are evaluated in a "hand powered" way. We plan to integrate the suitable software components coming from these packages to the Galaxy framework in a unified way (Galaxy or similar) to provide thorough tracking of numerical analyses done on the raw data and provenance of the secondary derived datasets.

Hardware resources are provided by Masaryk University, and they are fully integrated to the national e-infrastructure operated by CESNET. Temporary overflow (increased computing demand) is covered by general-purpose national e-infra resources.

Service Scope	**Service used**	**Limitation**	**EOSC service?**
AAI	Life-science AAI		yes
Workload Mng.	Galaxy/Slurm		possibly
Resource Mng.	Openstack		yes
Data Storage	Local filesystem based, Ceph		no
Monitoring	Icinga (local)		no
Computing Resources	MU + CESNET		yes
Storage Resources	MU + CESNET		yes

Table 36: Technical services identified for the UMSA Thematic Service.

A.8.2. Data and Workload Analysis

The application is in the development stage, details will be clarified gradually. We expect rather heterogeneous requirements (both high-throughput and large memory) in different stages of the processing pipeline. The expected application types are:

- *Parallel (shared memory multithreaded, distributed Memory, GPU-enabled)*

  Typically, the tools can leverage moderate multithreaded shared memory parallelism (16 cores); few of them use GPU.

- *High Throughput (sequential with many individual and independent jobs*

  There are two distinct patterns: when fresh data are processed, only few samples can be run in parallel; on the contrary, scans over historical data are embarrassingly high throughput (up to thousands samples in parallel)

- *Interactive*

Most of the processing can be run in batch mode, though a minority of tasks remains interactive

The applications are limited by both CPU and data (more or less balanced), although initial workflow components (baseline and noise removal and deconvolution) are memory demanding (up to terabyte) when applied on high-resolution data.

The vast majority of the used components are in production, the whole workflow is in early prototype stage.

Computing	Total CPU Consumption by time	50--100 CPU cores sustained; up to thousands cores when re-running analyses on historical data
	Number of individual jobs by time	Up to dozens of new samples per week (a single sample processing breaks up into few dozens of individual jobs); historic data analyses campaigns few times per year (thousands of jobs)
	RAM required by each single job.	Most stages only few GB, but some datasets may require up to hundreds GB in certain stages.
	Storage required by each single job.	Up to 100 GB.
	How does the application run?	Two major modes: - Processing of new data (fresh sample measurements), continuously - Re-running analyses on historical data, irregularly, few times per year
	Size of the input data per job	Up to 100 GB
	Size of the Output data per job:	Up to 100 GB
Storage	Permanent storage required	100-200 TB initially, will grow to PBs in several years
	Granularity (number of individual files)	Thousands of samples, typically hundreds of files per sample
	Access bandwidth required	10 Gbit/s
	Access pattern	Write once & read only for the primary data; R&W for the secondary data
Network	Does the application need to be accessed from outside?	Yes
	Does the application require downloading data from external sources? Yes	Yes

Table 37: Resource profile of UMSA.

A.9. MSWSS : Modelling Service for Water Supply System

A.9.1. Technical solution

This section describes the basic technical services needed to build the thematic service.

Service Scope	Service used	Limitation	Is an EOSC service?
AAI	EGI Check in		Yes
Workload Mng.	Slurm		No
Resource Mng.	IM, Ansible		Yes
Data Storage	MSWSS private storage + OneData	data confidentiality	Yes
Monitoring	Icinga		No
Other			
Computing Resources	IISAS Cloud, EOSC Synergy IaaS, EGI FedCloud	data confidentiality	Yes
Storage Resources	MSWSS private storage + OneData	data confidentiality	Yes

Table 38: Technical services identified for MSWSS Thematic Service.

A.9.2. Data and Workload Analysis

The MSWSS service will have these types of jobs:

- Sequential (single run)
  - One simulation (can include some of the post-processing tasks)
- High Throughput
  - 1 job consists of many (~ thousands) independent simulations
  - Simulations will be packed into blocks to decrease granularity and save bandwidth

The applications are mainly bound by CPU. The service also should deal with data confidentiality. The application is locally in production.

Computing	Total CPU Consumption by time	~20 CPU-hours per application run
	Number of individual jobs by time	on demand, one application run: ~20 jobs

	RAM required by each single job.	< 1 GB
	Storage required by each single job.	~ 1 GB
	How is the application run?	**periodically:** once per month per domain (area) **variable demand:** on request in critical situation, more simulations per week
	Size of the input data per job	~ 800 MB
	Size of the Output data per job:	~ 200 MB
Storage	Permanent storage required	2 TB
	Granularity (number of individual files)	~1000
	Access bandwidth required	low
	Access pattern	full files, R&W
Network	Does the application need to be accessed from outside?	Yes
	Does the application require downloading data from external sources?	Yes - pre-processing part of the MSWSS: ~100 MB, bandwidth is not critical - post-processing part of the MSWSS: ~ 1GB, modest bandwidth needed

Table 39: Resource profile of MSWSS.

A.10. O3AS: Ozone (O3) Assessment

A.10.1. Technical solution

Given the current conceptual development stage, we will describe the strategy and workflow, and will provide hard numbers while we develop the full workflow and service. A test service will be implemented on a cluster under Linux using virtualization. The full (test) workflow will be implemented on a dedicated VM. First, data will be accessed directly (from a file system mounted directly), later it will be available via a THREDDS data server. Data reduction steps will use CDOs

(https://code.mpimet.mpg.de/projects/cdo/) provided on the VM. Further calculations and front ends will be in python and a web interface will be created to allow easy access to the data, including a display function for a python based visualization. Monitoring of the VM will be aligned to the monitoring of other VMs running on the same cluster.
Preprocessing of data (creating selected time series that can be buffered) will be trialed. Based on the preprocessed (buffered) data different milestones can be quickly estimated (e.g. 1980 return dates – or return dates with respect to other years). The same will hold true for uncertainties.

Service Scope	**Service used**	**Limitation**	**EOSC service?**
AAI	Now: None Plan: EGI Checkin	Have a O3AS VO with sub groups	Yes
Workload Mng.	no specific need	--	
Resource Mng.	Now: Cloudify. Plan: IM / Other solution.	--	Hopefully
Data Storage	Now: Local at KIT Plan: WebDAV or OneData	--	No
Monitoring	No specific need, usage statistics should be nice	--	
Other			
Computing Resources	Now: Local HPC resources at KIT Plan: Access remote for data reduction prior to analysis		Currently there are no plans to offer cloud resources which are EOSC compatible.
Storage Resources	LSDF (KIT-SCC)		

Table 40: Technical services identified for OA3S Thematic Service.

A.10.2. Data and Workload Analysis

Computing	Total CPU Consumption by time	50-100 CPU cores, it depends on how many users will apply the service simultaneously. 168 CPU hours/week
	Number of individual jobs by time	>100 jobs/week (or more)
	RAM required by each single job.	16-20 GByte

	Storage required by each single job.	8-15 GByte per file
	How is the application run?	Via web-interface. Number crunching Python code in docker container.
	Size of the input data per job	16-20 GByte
	Size of the Output data per job:	8-10 GByte
Storage	Permanent storage required	300-500 TBytes
	Granularity (number of individual files)	~35000
	Access bandwidth required	1000 (or faster if possible)
	Access pattern	Full files, R&W
Network	Does the application need to be accessed from outside?	yes
	Does the application require downloading data from external sources?	- Yes - 8-10 GByte, 1000 MByte/sec

Table 41: Resource profile of O3AS.

B. Annex - DMPs

The second annex of this deliverable includes the Data Management Plans (DMPs) of the thematic services, summarized in section 5. The DMPs will evolve along the project time.

B.1. WorSiCa

Version	Date	Contributors
1.0	17/2/2020	Alberto Azevedo (LNEC)

B.1.1. Data summary

WorSiCa is a service that detects the coastline, coastal inundation areas and the limits of inland water bodies using remote sensing (satellite and UAVs) and in situ data (from field surveys). It is applicable to a range of purposes, from the determination of flooded areas (from rainfall, storms, hurricanes or tsunamis) to the detection of large water leaks in major water distribution networks. Therefore, the WorSiCa thematic service aims at integrating multiple-source remote sensing and in-situ data, integrated in a one-stop-shop service for remote sensing information without costs to all European public research groups. The integration of the WorSiCa service in the EOSC infrastructure will boost the usage of the service at a European level, taking advantage of the dissemination provided through EOSC channels, following FAIR data conformance directives and the availability of computational resources for its operation.

The WorSiCa service will process: Environmental data from satellites, UAVs, in situ data from field surveys and other open sources from EOSC-hub and European data Services. In summary, this service will use data that follows the Open Geospatial Consortium (OGC) standards (https://www.opengeospatial.org/standards). The data will be available in the users account, for a limited period of one month, due to storage limitations. After this period the users will be invited to download the data to their local servers. Some exceptions may occur if/when the user is a National Authority in a strategic field of research and offers their final products to the community in an open data approach. In these cases, the data will be stored in a permanent repository and will be freely available to the general public.

The input data used by the service will be delivered via several sources, such as: ESA/Copernicus, EMODnet, EOSC-hub OPENCoastS service, Pleiades satellite data provider, GEBCO bathymetry, Smith & Sandwell Topography and also in situ and UAVs data (provided by the users). The intermediate and final products of the WorSiCa service will be useful to researchers, civil protection authorities, environmental agencies and water management authorities. The end products will be applicable to a range of purposes, from the determination of flooded areas (from rainfall, storms, hurricanes or tsunamis) to the detection of large water leaks in major water distribution networks.

B.1.2. FAIR data

B.1.2.1 Making data findable, including provisions for metadata

WorSiCa will produce and reuse a variety of data types, from images to georeferenced information in GIS format. Metadata will be produced for all data using standards. Given that most data is of geographic nature, the INSPIRE directive (http://inspire-geoportal.ec.europa.eu/), the OGC Sensor Observation Service (SOS) Interface Standard (http://rd-alliance.github.io/metadata-directory/standards/observations-and-measurements.html) and ISO19115 (http://rd-alliance.github.io/metadata-directory/standards/iso-19115.html) standard appears to be the most appropriate.

Metadata creation is expected to be done using a metadata editor. The European Open Source Metadata Editor (EUOSME) is a web application to create INSPIRE-compliant metadata in any of 22 European languages. It is being developed and maintained by the Joint Research Centre as part of the EuroGEOSS project (www.eurogeoss.eu). This online editor is available at http://eenvplus.sinergis.it/euosmegwt/.

Consistency between metadata for similar data sets will be sought if standards are not available. Elements to be included in the metadata include a clear description of the data, the institution and person of contact responsible for the data creation, its format, creation date and possible modifications, data units and georeferencing (when applicable) and a number of keywords (metatags). The choice of adequate keywords will be included to promote and ease the discoverability of data. These keywords will include a number of fixed, common keywords in WorSiCa's scientific area and several new, free keywords that can help attract researchers from other areas to use and adapt WorSiCa's results to their scientific fields.

The open data provided by the users will be referenced with Digital Object Identifiers (DOI), assigned by the open data repository software service (e.g. Dataverse or ZENODO). This tool guarantees that all open data in WorSiCa will have persistent and unique identifiers. For consistency and promotion of data discovery, consistent naming conventions will also be used by the WorSiCa service.

Open access publications will also be sought, with direct links to the underlying data sets deposited in this open data repository.

In the WorSiCa service a succession of data sets will be produced, creating several databases of images at different stages of development and processing, from the input data to the intermediate and final processed products included in the end-user application. This sequence will be labelled with a unique identifier for each job submitted for a distinct Region of Interest (RoI), and final product type (e.g. coastline assessment, inland water bodies or water leak surveillance). A versioning policy will also be adopted and linked with detailed metadata and supporting documentation for each dataset produced by the WorSiCa service.

B.1.2.2 Making data openly accessible:

WorSiCa will create or reuse a variety of data sets, which have different natures and correspondingly distinct access privileges. Part of these privileges are set up accordingly with the data owner and acquisition policies.

A short overview on data access policies and availability is presented here. Regarding end-users' data, such as bathymetry data, UAVs field surveys or private Pleiades satellite images, these are categorized as classified and confidential, by default. Nevertheless, the field data sets may be categorized as open access data if the user possesses their ownership. The final products derived from classified data are also considered classified.

Classified data will be kept at the WorSiCa user account area repository, for a limited period of one month and can only be accessed by this user, fulfilling the policies of the data. After the one-month period the user must download/transfer the data from the WorSiCa account to a local server.

Regarding the processed products that are fully obtained from open access data, such as Sentinel imagery or the public bathymetry from EMODnet, the data will be categorized as open access data, by default. These data sets include mostly several databases of images at different stages of development and processing, ranging from the raw data derived from the hyperspectral to the processed products. The data can be accessed in the WorSiCa's web portal through the use of a web browser. In the WorSiCa service, the user has always the privilege to categorize the final products as classified, due, for instance, to classified Region of Interest.

Deposits in the open data repository will include the data, their metadata and their documentation. For most data sets, access is granted through generalized use software such as QGIS or similar.

B.1.2.3 Making data interoperable

All data developed in WorSiCa will be fully documented and accompanied with detailed metadata supported by a set of select keywords, to facilitate automatic discovery and integration of WorSiCa data for other purposes. Besides usual metadata fields, technical aspects such as units (complying with SI standards) and spatial and temporal references will be supplied. All data will be provided in generally used extensions, adopting well established formats (csv, shapefiles, netcdf, image formats, etc.) whenever possible which will also facilitate its use by other parties.

B.1.2.4 Increase data re-use (through clarifying licenses):

Since the usage of the WorSiCa service is made on-demand, the open data availability will occur as soon as the users authorize their publication, respecting the policies of the data used in the processing chain.

The usefulness of the data for third parties is closely linked to the perception of quality and robustness of the available data. Although the service provider is not responsible for the quality of the final products, the methodology behind WorSiCa will be clearly provided to the users (through user's manual and publications) to enhance confidence and re-usage of the outcomes.

B.1.3. Allocation of resources

In WorSiCa there is a considerable amount of data that will be managed by the service in the IberGRID infrastructure. This will be the principal storage hub for WorSiCa.
A complementary storage will be created with a repository service like Dataverse (https://dataverse.org) or ZENODO (https://zenodo.org), for the storage of the open data generated by the WorSiCa service and shared by the users, therefore ensuring data availability, backup and versioning.
Publications and technical reports featuring the data will be produced by the WorSiCa's development team and the users of the service. These products will be made available through open access (using open access journals or journals selected for a short embargo period). This channel will provide a long-term availability of data and data analysis.
The WorSiCa DMP will be updated throughout the EOSC-Synergy project.

B.1.4. Data security

Open data security will be addressed in WorSiCa taking advantage of Dataverse's or ZENODO's services of secure storage, backup and preservation and protected transfer mechanisms.
Regarding the confidential data, data will be housed on servers under direct management of the institution's personnel to be installed in already provisioned data centers. These data centers are expected to be equipped with various features ranging from secure physical access, air conditioning, generators and fire extinguishing measures. Typically, hardware/electricity failure are addressed with redundant hardware and generators.
Access to data under different permission conditions (read-only, read-write, etc.) are granted to users and authorized computers by project managers or to whomever this task is delegated, according to a well-defined protocol. Taking in account the size of the data at stake that requires regular backup (be it either for security versus a hardware failure or for archival purposes), a sequence of regular full backups, differential backups and incremental backups on an increasingly frequent basis are envisaged and following already installed procedures. The physical media used to store the data will be maintained in secure locations. Access to these backups is limited to the personnel authorized to use the backup system, and as a general rule, not authorized for external sources.
All data transfers should be encrypted to render all stolen/lost data useless. Encryption methods are to be specified at a later date.

B.1.5. Ethical aspects

The WorSiCa partners are to comply with the ethical principles as set out in Article 34 of the Grant Agreement, which states that all activities must be carried out in compliance with:

(a) ethical principles (including the highest standards of research integrity) and
(b) applicable international, EU and national law.

Activities raising ethical issues must comply with the 'ethics requirements' set out as deliverables in Annex 1 of the Grant Agreement.
Activities raising ethical issues must comply with the "ethics requirements" set out in Annex 1 of the Grant Agreement.

B.2. G-Core

Version	Date	Contributors
1.0	13/2/2020	Juan Sánchez-Ferrero de Pablo (INDRA)

B.2.1. Data summary

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation**. Gcore is a Cloud Data Processing Service system that allows performing the main activities associated to manage the Payload Data Ground Segment (PDGS) of a satellite's ground segment associated mission, incorporating the new cloud paradigm to reduce the deployment cost and to take advantage of their capacities to increase and decrease the resources automatically. As GCore pretends to be as versatile as possible, it is prepared to deal with a great variety of data. In that sense the main data handled by the system would be the data associated for each particular satellite mission. Typically, these data sets (raw data) are based on the sensors on board the satellites producing products with metadata associated. These products will depend on the mission. For example, Earth Observation missions will be based on optical or radar imagery with its associated metadata. In addition to the reception of this data, the system also will produce data after performing a processing of these products in order to obtain higher value products (L1, L2 data) to be delivered between the users and customers of the system.
- **Explain the relation to the objectives of the project.** The objective of the adaptation of the thematic service is to explore the sustainability of the EOSC services exposed through the creation of added-value products through the integration of G-Core as a data manager. In the other hand G-Core can also provide new functionalities, standards (Inspire directive based metadata) and interfaces (OGC based) used in the Earth Observation fields as well as new cloud capabilities to the rest of thematic studies that uses the EOSC platform in order to enrich the cross data distribution between different scientific fields.
- **Specify the types and formats of data generated/collected.** Satellite raw data are generally received as ISPs (Instrument source packets) binary files for being processed by the processors. These data once processed are stored as tar.gz files. These files are very dependent of the mission but usually have a defined structure that typically can be: A main folder with the product's name with several subfolders for the different components of the product as Annotation, imagedata, preview, auxraster, support, calibration, etc. These subfolders contain the files needed to treat the product, for example annotation will have the xml metadata file with all the information related to the product, imagedata will have the imagery in tiff, geoTIFF or JPG2000 format. Preview will have small quicklooks of the image in jpg or tiff and xml files, etc.
- **Specify if existing data is being re-used (if any)**. The system can re-use previous satellite data products, for example the Sentinel data is available in ESA data hubs and DIAS platforms and can be used and distributed free of charge between the scientific application that will need them.

Another example is for re-processing campaigns in case of having a new version of the processor that will need to process the entire data gathered up to date for a particular mission.

- **Specify the origin of the data.** They can be either satellite raw data or existing data of scientific applications to be managed by Gcore.
- **State the expected size of the data (if known).** In principle Gcore is prepared to deal with the different size of products that can go from a few Kbytes or Mbytes from xml or auxiliary data to tens of Gbytes of final product imagery which are the usual range for EO products. Example: ISP: 360Mbytes – 1700Mbytes, L1B processed: 745Mbytes-7500Mbytes for single products.
- **Outline the data utility: to whom will it be useful.** It is difficult to summarize because satellite data can be useful for a wide range of disciplines and areas. We have in mind two goals: Popularize the satellite imagery data and allow crossing data with other disciplines in order to achieve cross-fertilization in the investigation field to produce better results and services.

B.2.2. FAIR data

B.2.2.1 Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision).** The different products managed by Gcore are stored in a catalogue. The metadata is extracted from the products and is used to fill the catalogue of products. The discovery of products is available with the Elasticsearch engine which is a distributed database management system that allows to perform fast access to the catalogue, in addition there are also an CSW ebRIM tailored implementation to publish the catalogue and perform queries.
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?** All missions and all products have unique identifiers which allow them to identify the product at any moment as the system is data driven no Id repetitions are allowed. There is no DOI mechanism implemented.
- **Outline naming conventions used.** Every mission or product has its own naming convention defined to guarantee the uniqueness of file naming.
- **Outline the approach towards search keywords.** Search will be based in any metadata value, as an example the data available for EO missions are: Satellite, Sensing time, AOI (Area Of Interest), Sensor type, Imaging Mode, Processing mode, Polarization, Incidence angle, Resolution mode, etc.
- **Outline the approach for clear versioning.** The naming convention includes the possibility to produce different versions of one product or file. This case is typical for reprocessing of EO products. The resulting products are produced with the same name but with a tail version in the name of the file denoting the version of the product. This way the uniqueness of the file naming keeps guaranteed in the system.
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how.** Inspire directives are usually followed up by EO missions, OGC based metadata also is widely used.

B.2.2.2 Making data openly accessible:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so.** It will depend on the satellite missions, some of them have free access policies like Copernicus (sentinel series) or ESA scientific missions like SMOS and others are commercial services like Pleiades. Gcore as data manager will implement the policy of the data owner and will not impose any restriction to the original data.
- **Specify how the data will be made available.** Gcore has a specific component for dissemination. This component is configurable and can work in pull or push mode. Usually the products once are ready to be delivered are located in a temporal repository to be retrieved by customers.
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?** There are two methods to access the catalogue. The former consists of access through a viewer tool to perform the request to the catalogue in order to visualize the products, the latter is an electronic interface based in HMA CSW ebRIM taylored. This interface allows you to send HMA Catalogue interface requests using the HMA Catalogue EO Products Extension Package for ebRIM (ISO/TS 15000-3) Profile of CSW 2.0 Service Protocol. The purpose of this interface is to provide the capacity to browse Products and the retrieval of the Past Data Request from catalogue using SOAP instead HMI. HMA Catalogue Interface shall provide the next operations:
    - GetCapabilities
    - GetRecords
    - GetRecordsById
- **Specify where the data and associated metadata, documentation and code are deposited.** Data and metadata can be stored either in EGI DataHub service or in any external cloud provider decided by the customer/user. The documentation will be held in the Help services available or via the application's help functionality. The source code is not available.
- **Specify how access will be provided in case there are any restrictions.** For open and free data there are no restrictions, previously the user needed to be registered and validated to access the data.

B.2.2.3 Making data interoperable:

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.** The GCore provides interfaces for commercial partners based on tailored versions of the HMA protocol and WFS interface. These interfaces define an interoperable method to access the system capabilities to perform the tasking activities and ordering services for EO missions. The HMA is a technical proposal promoted by international agencies in the scope of EO data to define a common standard to codify the metadata information of EO products as geographic features encoded in OGC Geographic Markup Language. The main reason to adopt the gml notation instead of other type of structure for encoding this type of products is the flexibility and versatility of this language to codify the specific parameters of EO products according to ISO 19115.

All metadata elements common to all Earth Observation products were defined within an Earth Observation Product (eop) GML application schema, formerly known as HMA schema. Specific metadata elements for optical, radar and atmospheric products, were assigned to three specific application schemas deriving (respectively opt, sar and atm) from the base eop schema. For products of specific missions requiring further metadata elements for their descriptions, it is possible to define a specific application schema deriving from one of the thematic application schemas.

Since the initial work on the GML Application Schema for EO Products in 2006, the base GML 3.1.1 specification of which [OGC 06-080r4] is an application schema has been superseded by a newer version. GML 3.2.1 [OGC 07-036] is now the official OGC GML Implementation Specification since July 2007. It was therefore logical to align this new version of the specification with GML 3.2.1 which is also used within O&M and WCS 2.0.

On the other hand, the WFS interface provides a standard interface supported by OGC to publish and search geographic features stored in public catalogues to be discovered and binded by external parties and resources. These features are based on metadata that contain the descriptive information of geospatial data.

The OGC Web Feature interface also provides the technical specification to develop frameworks to define different application profiles needed to publish and access digital features and acts as catalogues of metadata for geospatial data, services, and associated resource information.

- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?** Gml notation for product's encoding to codify the specific parameters of EO products according to ISO 19115. Additional mapping to other scientific fields needs to be analyzed case by case.

B.2.2.4 Increase data re-use (through clarifying licenses):

- **Specify how the data will be licenced to permit the widest reuse possible.** Since Gcore is not a data producer but a data manager, the data will be subject to the licences and policies of the owners. Gcore will not add any restriction in that sense.
- **Specify when the data will be made available for reuse. If applicable, specify why and for what period a data embargo is needed.** Since Gcore is not a data producer but a data manager the owner of data will decide if an embargo period is applicable. The Gcore is ready to hide the products in the catalogue for a defined period in case of necessity.
- **Specify whether the data produced and/or used in the project is usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.** Same answers that above.
- **Describe data quality assurance processes.** Gcore has a schema checking for interchanged files validation, but as a data manager there is no specific quality control implemented.
- **Specify the length of time for which the data will remain re-usable.** It is a decision of the data owner to decide how much time the data are kept in the archive.

B.2.3. Allocation of resources

Explain the allocation of resources, addressing the following issues:

- **Estimate the costs for making your data FAIR. Describe how you intend to cover these costs.** Gcore since its origin treats data to be findable, accessible, interoperable and reusable as recommended by the FAIR guiding principles. As data manager Gcore will use data following these principles and the standards and specifications previously mentioned. Gcore does not envisage extra development to adapt the data unless some minor modification could be needed to address very specific issues.
- **Clearly identify responsibilities for data management in your project.** In principle as data manager the main responsibility falls in the processing flow, the catalogue, the archive and the dissemination of the product being some of the responsibilities shared with the cloud provider and the data owner.
- **Describe costs and potential value of long term preservation.** This will depend on the project that will use Gcore as a service for either using it as a PDGS or as a processing and data management service.

B.2.4. Data security

Address data recovery as well as secure storage and transfer of sensitive data. Data recovery and secure storage shall be shared with the cloud provider. As previously mentioned the Gcore is open to any type of data, this means that the data can be open or sensitive it will depend on the data owner. In case of sensitivity data, special measures for data storing must be agreed with the cloud provider in order to guarantee the isolation and limiting the data access.

B.2.5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former. Gcore itself has no data that can represent ethical or security issues. This aspect will need to be analyzed every time a new user wants to make use of Gcore with its own collection of data.

B.3. SAPS

Version	Date	Contributors
1.0	17/2/2020	Thiago Emmanuel Da Silva (UFCG), Francisco Brasileiro (UFCG), Amanda Calatrava (UPV), Ignacio Blanquer (UPV)

B.3.1. Data summary

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation:** SAPS is a service to compute the Surface Energy Balance Algorithm for Land (SEBAL) and similar information for estimating the evolution of forest masses and crop. The data generated by the service will provide wider access to knowledge on the impact of human and environmental actions on vegetations, leading better forest management and analysis of risks.
- **Explain the relation to the objectives of the project**: SAPS is one of the Thematic Services of the EOSC-Synergy project. The aim is to promote EOSC adoption by the research communities, represented by the Tematic services by increasing knowledge on common interfaces, standards and best practices. This will be supported by an expansion of the capacity through the federation of compute, storage and data resources aligned with the EOSC and FAIR policies and practices.
- **Specify the types and formats of data generated/collected**: an execution of a job generates a bunch of files, one of them being the description of the generated files. We also keep this description into a service catalog.
- **Specify if existing data is being re-used (if any)**: The output data is potentially useful for a long time; whether the decision of keeping the data or not for a long time is a decision of the service provider, though.
- **Specify the origin of the data**: SAPS uses the data from the LANDSAT data repository (http://landsat.gsfc.nasa.gov/). It also analyzes Meteorological information provided by the National Centers for Environment Information (NCEI - ftp://ftp.ncdc.noaa.gov/pub/data/noaa/), and elevation data provided by the Consortium Spatial Information (CS - http://srtm.csi.cgiar.org/).
- **State the expected size of the data (if known):** Each job of the SAPS pipeline produces 4GB of output data. The experiment we are proposing to do will be composed of around 27.000 jobs. Therefore, we expect to have around 108 Terabytes of data.
- **Outline the data utility**: Researchers in Agriculture Engineering and Environment.

B.3.2. FAIR data

B.3.2.1 Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision)**: The metadata describing the data generated by SAPS is stored in a catalog. The catalog content is exposed through a REST API.

Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?: All the data generated by SAPS are grouped together based on a unique identifier of the job that generated the data. SAPS does not make use of DOI.

- **Outline naming conventions used**: The algorithms available in the SAPS platform are required to generate a special file that describes all the data generated by the algorithm and how they are named.
- **Outline the approach towards search keyword**: Search on the catalog is based on the region of the globe, the date period (of satellite capture) and the algorithms used to process the data.
- **Outline the approach for clear versioning**: SAPS does not version the generated data.
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**: SAPS follows the PROV standard to describe the metadata stored in the catalog.

B.3.2.2 Making data openly accessible:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so**: all data generated is available to all registered users. There is no license and registration in the SAPS service is open and free.
- **Specify how the data will be made available**: Through the SAPS web service. SAPS will reply with URLs pointing to the data that match the query.
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?** The users only need a web browser to connect with the SAPS portal and access the data. All the SAPS code is currently open source and available at: https://github.com/ufcg-lsd/saps-engine
- **Specify where the data and associated metadata, documentation and code are deposited**: The data is stored in a permanent storage managed by Openstack Swift.
- **Specify how access will be provided in case there are any restrictions**: Users just need to be registered in the SAPS service to access the data. The registration is free and open to the community.

B.3.2.3 Making data interoperable:

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability:** The SAPS platform processes landsat satellite imagery. As such, the algorithms available in the SAPS platform typically adopt the standard satellite data formats, including the TIFF and the NetCDF ones.
- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?:** SAPS adopts the PROV standard format to describe the metadata content. A common vocabulary could be promoted but is not a requirement.

B.3.2.4 Increase data re-use (through clarifying licenses):

- **Specify how the data will be licenced to permit the widest reuse possible**: Currently, produced data has no license. However, we plan to add a license to allow sharing the data among users and protecting it at the same time. Licenses like Creative Commons (CC), Open Data Commons (ODP) or Open Government Licence (OGL) will be explored.
- **Specify when the data will be made available for reuse. If applicable, specify why and for what period a data embargo is needed**: As soon as the execution of the workflow finishes, the data will be available for SAPS users. No embargo is required.
- **Describe data quality assurance processes**: SAPS has no automatic QA process implemented. This can be done by the users when they create customized pipelines (the last phase of the pipeline could perform some automatic checking of the consistency/accuracy of the output generated).
- **Specify the length of time for which the data will remain re-usable**: whether the data will be kept or not for a long time is a decision of the service provider.

B.3.3. Allocation of resources

Explain the allocation of resources, addressing the following issues:

- **Estimate the costs for making your data FAIR. Describe how you intend to cover these costs**. The cost during the project lifetime will be jointly covered by UPV and UFCG, exploring ways to fund the sustainability of the services and the contention of costs.
- **Clearly identify responsibilities for data management in your project**. Data management will be the responsibility of the site that hosts the data and the services (UPV and UFCG).
- **Describe costs and potential value of long term preservation**. Data is valid for a long time. As processing required to obtain the results is lengthy and intense, data will remain valuable in the medium-term. The project will study the use of long-term preservation repositories for such data.

B.3.4. Data security

Address data recovery as well as secure storage and transfer of sensitive data: SAPS does not work with sensitive data, so no special need of anonymization and encryption of the data is required. Regarding data recovery, this relies on the service provider.

B.3.5. Ethical aspects

No ethical issues apply to the data.

B.4. OpenEBench

Version	Date	Contributors
1.0	2020-02-16	Laia Codó, Salvador Capella-Gutierrez, José Mª Fernández, Josep Ll. Gelpi

The dependence of the scientific advance on research software is increasing in all science fields. Notably in biology, where the availability of growing amounts of data coming from large scale genomics projects has put an extra concern in the possibility of properly analyzing such data, and hence assuring the outcomes of such projects. Bioinformatics as a science has become a need at all levels of biology. It is no longer a private space where some specialized researchers develop and test new methodologies for the sake of their own scientific objectives. Bioinformatic methods and tools have now to be consumed by the whole of the biological community. This puts an extra challenge in the development of research software. Bioinformaticians should prepare software for the use of non experts, and have to compete in a continuously evolving market of alternative options, proving with objective metrics that the software is usable, efficient, and gives the adequate answers. Benchmarking has been a traditional activity in bioinformatics, although it has been mostly conducted by scientific communities, for internal consumption and seldom considered by final users of the software.

With the advent of different personalized medicine initiatives, there is an emerging need to guarantee, and to a certain extent to certify, that analytical workflows used routinely in the clinical practice are complaint with the highest standards, implement state-of-the-art technologies and consistently process input data as expected. Thus, there is a clear need of establishing standards, relevant scientific challenges and meaningful metrics by knowledgeable scientific communities. However, those efforts should be complemented by a stable platform which can support these activities, provide a reference place for different stakeholders and give a general overview on how tools and workflows, scientific challenges, metrics and data sets evolve over time.

In this context, the need for an open platform around benchmarking has become evident. **OpenEBench** (https://openebench.bsc.es), one of the main H2020 ELIXIR-EXCELERATE outcomes seeks to fill in this gap and three different but yet complementary levels of benchmarking: i) scientific benchmarking related to the scientific quality of bioinformatics tools and workflows; ii) technical monitoring related to software quality; and iii) performance benchmarking regarding the usability and efficiency of the technical deployment of bioinformatics tools, servers and/or workflows. Indeed, OpenEBench aims to provide information for i) end-users, deciding which resource is the most appropriate for their problem at hand, ii) software developers, seeking for accepted best practices in research software, and testing their own tools against the accepted and/or possibly competing alternatives, iii) infrastructure providers, seeking to design an adequate provision of tools, servers and/or workflows, and iv) funders, requiring an overview of a given field, and checking the outcome of funded activities. A number of other initiatives do exist within and outside ELIXIR that clearly intersects of OpenEBench aims. In particular, tool's registries, mainly bio.tools registry (https://bio.tools), aggregated tools platforms like BioConda or/and Galaxy tool-shed, or software deployment platforms like BioContainers.

B.4.1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

OpenEBench is designed as an information Hub, where data is being collected from different sources, processed, and redistributed back for the use of interconnected platforms and scientific communities. The main objective of OpenEBench within EOSC-Synergy is to become a reference point for communities within the Life Sciences interested in pushing forward scientific benchmarking activities. Indeed, the OpenEBench central data repository is populated with well-organized, structured and validated data sets associated with the performance of bioinformatics software resources for one or more benchmarking challenges. Apart from the datasets brought by the scientific communities to OpenEBench, the OpenEBench Virtual Research Environment (VRE) is a second source of benchmarking data generation within the platform. OpenEBench VRE offers an online workbench to software developers for evaluating the scientific performance of their own methods under controlled circumstances using datasets and metrics defined for each community.

What types and formats of data will the project generate/collect?

Data types and formats of the data deposited at OpenEBench are specific to each scientific provider community. OpenEBench will maintain those data management criteria to be able to assure its interoperability with the participating communities. However, in an effort to standardize the benchmarking process per se, we have developed a refined data-model to reflect the process itself and allow scientists to refer to a particular step and/or data set in a defined way. Figure B.4.1 depicts the workflow for a single Benchmarking Event. Participants represent those systems e.g. individual tools, analytical workflows, web-servers, taking part of a specific benchmark event. Details of the OpenEBench data model are available at the GitHub Repository (https://github.com/inab/benchmarking-data-model). A detailed explanation of created data sets types follows:

- **Public Reference data sets**. They are a widespread, publicly available and well characterized data set which can be used by developers and/or interested users to gather performance data of their systems in a controlled set-up. Scientific communities tend to make available Public Reference data to facilitate the engagement of participants within the challenges at hand. These data sets could comprise data from previous benchmarking editions but it is highly dependent on the community and the scientific problem at hand.
- **Input data sets**. Represent the data sets to be processed as input by participants in the benchmarking activities. Those data sets can be publicly available for download at specific repositories e.g. UniProtKB specific reference proteome sets for the Quest for Orthologs participants; and/or can be submitted automatically by benchmarking platform e.g. CAMEO, to participants web-servers. Input data sets should follow at least the same data formats as the Public Reference data sets, and should provide enough metadata describing the data sets to facilitate reproducibility, data provenance and, potentially, the evolution of participants across different benchmarking challenges editions with different input data sets of varying degrees of complexity.
- **Participant data sets**. These data sets represent the data e.g. predictions, produced by participants given a specific Input data set associated to specific benchmarking activities. Depending on the level of automation, participant data sets can be submitted manually e.g. uploaded to a server,

and/or automatically e.g. response via APIs implemented in systems like BeCalm. Unless previously agreed, participant data sets are often kept private to participants and/or communities. It would be recommendable that participant data sets which are part of scientific benchmarking publications should be made available for reproducibility purposes, data reuse in downstream analysis and/or further meta-analysis.

- **Metrics Reference data sets**. These data sets contain data used to evaluate the benchmarking process, i.e. the "true" responses to the challenges. These data sets are often kept private by benchmarking events organizers while a challenge is active. This standard practice prevents participants from adjusting their systems to have the best performance for very specific data sets (overfitting). Overfitting may render systems useless and not-fit-to-purpose and, therefore, it is highly discouraged. Depending on the nature of the Metrics Reference data sets, those can be either "Gold data sets" or "Silver data sets". It is not uncommon to have both types of data sets as part of a Benchmarking event. When available, Golden data is desirable because it represents the ultimate data that any system should aim to produce. For instance, in the case of Protein Structure Predictions the experimental data deposited in the Protein Data Bank (PDB) is considered to be the "Gold data" for the benchmarking activities carried out by communities such as CAMEO, CASP, and CAPRI. In the absence of a gold standard, benchmarking efforts have to resort to "Silver data". For instance, synthetic and/or simulated datasets generated in silico following previous experiences[13] or with data generated using unsupervised learning approaches, based on the consensus among different —i.e. algorithmically independent — methods[14]. For the latter, naive methods e.g. Bayesian networks, can provide a baseline allowing assessors to measure relative performance between methods with, on average, moderate to good accuracy. Such consensus data is referred to as "Silver data". However, data from silver standards should be used with caution as it needs to be revised regularly to adequately evaluate new developments in the field. Often Metrics Reference data sets become public e.g. Public Reference data sets, once a given challenge has concluded because of its intrinsic value to address valuable scientific challenges.
- **Assessment data sets**. These data sets are produced after applying specific metrics e.g. Q50, to participants data sets while considering metrics reference data sets. Assessment data sets establishes how close or far are participants from the expected results. Often preliminary assessment data sets tend to be private to each participant e.g. understanding the initial characteristics of the platforms and/or metrics reference data sets nature; while final assessment data sets tend to be shared among benchmarking participants before the challenge ends, and made public once the events end. Even when participant data sets are not available, assessment data sets can be very useful to measure the performance evolution of different systems versions for the same challenge and/or the complexity of different reference metrics data sets for the same system. Ideally, assessment data sets would allow to track the evolution of both reference metrics data sets and systems versions. However, it would be nearly impossible to deconvolute the impact of each variable into the final results.

[13]Hatem, A., Bozdağ, D., Toland, A. E., & Çatalyürek, Ü. V. (2013). Benchmarking short sequence mapping tools. BMC Bioinformatics, 14(1), 184. https://doi.org/10.1186/1471-2105-14-184

[14] Elsik, C. G., Mackey, A. J., Reese, J. T., Milshina, N. V., Roos, D. S., & Weinstock, G. M. (2007). Genome Biology, 8(1), R13. https://doi.org/10.1186/gb-2007-8-1-r13

- **Challenge data sets**. These data sets are considered metadata sets grouping either i) assessment data sets from different participants for the same reference metrics data set and applied metrics, ii) assessment data sets from the same participant but for different reference metrics data sets and/or applied metrics in the same benchmarking event, or iii) the grouping of the assessment data sets from the same participant and the same applied metrics across different benchmarking events. Challenge data sets are the foundations of the community-led scientific benchmarking activities as they offer an unified framework to compare participants performance among themselves for a specific scientific challenge and/or the evolution of individual participants along time. Challenge data sets allow data bundling and are the ones consumed by experts and non-experts for taking decisions on what systems to use for their own scientific problems. Challenge data sets can be directly offered at OpenEBench using available views e.g. experts and non-experts data views; and/or using available APIs. Those data sets due to their own nature would be mostly public although they might remain private to scientific communities and/or benchmarking participants while challenges remain open.

Each Benchmarking event can be represented by a data flow composed by these six different data types, as illustrated in figure 2. In the case of continuous benchmarking systems, the red arrow at figure 2 indicates the start of the subsequent cycles which often tend to keep the same metrics and change the Reference Metrics data sets e.g. CAMEO (https://www.cameo3d.org/).

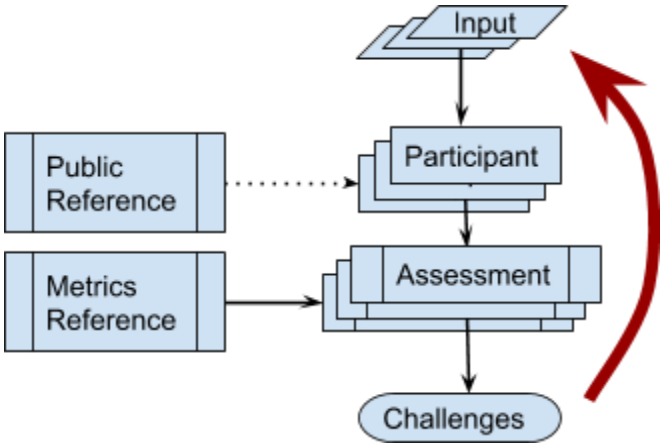

Figure B.4.1. OpenEBench definition of datasets and how they relate to each other.

Will you re-use any existing data and how?

There are few cases where datasets are generated *ex professo* for supporting benchmarking activities (e.g. CAMI, https://data.cami-challenge.org/). In most of the cases, Scientific Communities decide to use high-quality datasets generated for other purposes for carrying on benchmarking activities. Often, the datasets assembled by scientific communities are composed of data representing different aspects of the current challenges within a given scientific domain.

What is the origin of the data?

Generally speaking, datasets are generated within public and/or private projects as part of the normal scientific activities within each community. Some communities may re-use potentially sensitive research data for benchmarking process.

What is the expected size of the data?

Datasets are highly dependent on each community. However, datasets generated following the OpenEBench data model amounts for a few kilobytes to megabytes.

To whom might it be useful ('data utility')?

As stated before, datasets are not generated in many cases for supporting benchmarking activities. But datasets generated for other purposes are useful to organize benchmarking activities within scientific communities across Life Sciences. Aggregated data organized and/or generated within OpenEBench is useful to users in order to make informed decisions regarding which one is the best resource for a specific scientific question and/or to identify potential areas of improvement for software developers.

B.4.2. FAIR data

B.4.2.1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

OpenEBench aims to be a transversal infrastructure able to promote the generation, exchange and sharing of benchmarking data across multiple scientific communities. As such, building a comprehensive and mature common data model focused on technical and scientific benchmarking terminology has been a priority. The model has established an inclusive and comprehensive metadata framework for the platform's data services. Implemented as a set of JSON schemas[15], the OpenEBench benchmarking data model organizes and annotates OpenEBench benchmarking concepts. The model is widely adopted by OpenEBench services, who ensure that all received and generated data is accompanied by the right set of metadata, either when pushed into the central OpenEBench repositories via REST APIs, or when generated on the OpenBench-VRE framework. Datasets are localable by means of standard identification mechanisms like unique identifiers under versioned records, which in turn, might agglutinate annotations pointing to other resources, always referred via unique resource links (URLs/URIs).

OpenEBench plans to offer long-term persistent identifiers, i.e. Digital Object Identifiers, via sustainable data archives infrastructures like Zenodo[16] or EUDAT[17]. Internal OpenEBench identifiers will be made findable through the identifiers.org service.

What naming conventions do you follow?

Identifiers across the OpenEBench data model follows the recommendations made by the community towards findable IDs that require them to be unique, persistent and permanent[18].

Will search keywords be provided that optimize possibilities for re-use?

The complete set of metadata provides categorized search keywords to promote data re-use across not only different benchmarking challenges and events, but also across different scientific communities. Services provide complete querying systems on such metadata.

Do you provide clear version numbers?

[15] https://openebench.bsc.es/docs/oeb/benchmarking-data-model
[16] Zenodo, catch-all repository for EC funded research supported by OpenAIRE. https://zenodo.org/
[17] EUDAT, Research Data Services, Expertise & Technology Solutions. https://www.eudat.eu/
[18] Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. McMurry JA *et al*. PLOS Biology 15(6):e2001414. doi:10.1371/journal.pbio.2001414.

Versioning is an integral component of OpenEBench data management. In the case of Benchmarking events and challenges, all data is maintained related to the organized events. Datasets not related to specific events are clearly tagged with the appropriate versions and maintained. Software versions for benchmarked tools are clearly stated and integrated in the benchmarking results.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

OpenEBench data model provides descriptive and operational metadata on benchmarking datasets and tools to guarantee findability, provenance and reproducibility. The model has been developed within the project in the absence of standard metadata in the field. In those cases where the annotated data is not specifically from the benchmarking domain, standard and well-known ontologies are being adopted to categorize the files. For instance, an extension of EDAM[19] ontological terms are in use to define participant's datasets in OpenEBench-VRE. A benchmarking specific ontology is being developed.

B.4.2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

OpenEBench platform aims to be a central platform not only to generate, but to publish and distribute benchmarking data across the scientific community. To this end, a set of microservices are publicly offered as REST APIs to retrieve data from the major OpenEBench repositories.

Data Retrieval APIs	Endpoint	Source code
OpenEBench Tools Monitoring	https://openebench.bsc.es/monitor/	https://github.com/inab/elixibilitas
OpenEBench Scientific	https://openebench.bsc.es/openebench/rest/breed/	https://github.com/inab/oeb-benchmarking-api/tree/master/api-rest/java

Although benchmarking datasets are generally in the public domain, some communities do re-use potentially sensitive data either from personal and/or commercial origin. In such cases, the necessary agreements with data providers are met. Access to OpenEBench is generally authenticated (although anonymous users can be created). In those conditions data and tools access can be restricted as required. OpenEBench will not provide data access credentials. Instead, we will honor the agreements between data users and providers.

What methods or software tools are needed to access the data?

OpenEBench data is always accessible on the internet via HTTP(s), either via RESTful APIs, or using the web applications and widgets fed by them.

Is documentation about the software needed to access the data included?

[19] EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Ison, J. *et al*. *Bioinformatics*, *29*(10), 1325-1332. https://doi.org/10.1093/bioinformatics/btt113

While following the general recommendation of openAPI RESTful resources, most API-based operations result intuitive and self-explanatory. However, endpoint's specifications and data models in use are well documented to better understand service capabilities.

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Benchmarking data is being processed and deposited at the ELIXIR-ES private cloud hosted at the Barcelona Supercomputing Center, where as above detailed, several mechanisms have been put in place to openly access it.

On the other side, raw benchmarking datasets belong already to the public domain, typically maintained by the different scientific communities.

Have you explored appropriate arrangements with the identified repository?

Yes. OpenEBench namespace is being registered at identifiers.org as a mechanism to resolve OpenEBench IDs when used across the Internet.

B.4.3. Allocation of resources

OpenEBench-VRE is running in a series of on-premise private cloud infrastructures based at ELIXIR partners institutions. Physical compute clusters hosting the cloud infrastructures are physically protected and follow the security standards of hosting institutions. Direct access to the cluster is reserved to technical staff. Data transmission necessary for the internal activity of OpenEBench-VRE use in all cases secure protocols (https, sftp, and similar).

B.4.4. Data security

OpenEBench-VRE uses Keycloak software (https://www.keycloak.org/) to manage the authentication layer based on OpenID Connect protocols. For registered users access is protected by a username and password combination, issued by the central Authentication service. Alternatively, the task might be delegated to existing trusted identity providers (IdPs), currently Google OpenID Connect and ELIXIR Authentication and Authorization Infrastructure, (AAI), also named EOSC Life Identifier. Additional identity providers may be added in the future including GA4GH AAI profile and GA4GH passport claims.

OpenEBench-VRE portal supports HTTPS access, which ensures encryption of all transferred data. Authentication protocols are applied transparently to all data transactions between the user and the portal, either interactively, or through REST based APIs to assure privacy.

B.4.5. Ethical aspects

In accordance with Organic Law 15/1999 of December 13 on Protection of Personal Data, 3/2018 of December 5 of Personal Data Protection and Guarantee of Digital Rights, and General Data Protection Regulation 2016/679, personal data provided in the registration will be incorporated into a file property of the ELIXIR-ES/INB Hub Consortium, located in C/ Jordi Girona n.31, 08034 Barcelona (Spain) and will be treated in a confidential way. Use of this file is restricted to the needs of OpenEBench-VRE and its contents will not be shared with third parties. We remind you that at any time you may exert your rights of access, modification, rectification or removal, the limitation of the treatment or opposition to it, as

well as the right to data portability by contacting ELIXIR-ES/INB Hub Coordinator at the Barcelona Supercomputing Center in writing (BSC, C/ Jordi Girona n.31, 08034 Barcelona (Spain)) or by email inb.hub@bsc.es.

B.4.6. Further support in developing your DMP

The team at OpenEBench is part of the newly established H2020 ELIXIR Converge[20] project (as february 1st, 2020) that is dedicated to develop specific tooling for Life Sciences data management plan through use-cases.

[20] https://elixir-europe.org/about-us/how-funded/eu-projects/converge

B.5. Scipion - Instruct-ERIC Data Management Policy

Version	Date	Contributors
2.1	2/11/2018	José María Carazo (CNB)

1. Introduction

The purpose of this policy is to provide Users conducting Instruct-ERIC Access projects with information and guidance on Experimental Data ownership, storage, access and management and to ensure that Experimental Data is managed and used in ways that maximises public benefit following FAIR principles (Findability, Accessibility, Interoperability, and Reusability). This policy should be read in conjunction with the Instruct-ERIC Statutes (2017/C 230/01).

The Structural Biology community has had a historical commitment to make the processed data and the structural models available to the public via the PDB, the oldest biological data archive. In a continuation of this tradition, and in line with the perspective of the European Commission that data from publicly funded research projects is public data [COM(2011) 882 final], Instruct-ERIC encourages experimental data sharing and reuse.

2. Policy applicability

This policy applies to Users of facilities at Instruct-ERIC Centres, which conduct Instruct-ERIC Access projects and produce Experimental Data. Experimental Data arising from Proprietary Research is not covered by this policy and is subject to separate contractual arrangements.

3. Policy responsibility

The Instruct-ERIC Council has overall responsibility for this policy. Any queries or suggestions relating to this policy should be sent to the Instruct-ERIC-ERIC Director.

4. Definitions

Terms and phrases in this policy shall have the meanings ascribed to them below.

“Access Proposal (or Access project)”: A research proposal describing a limited work programme that requests access to one or more Instruct-ERIC infrastructure facilities through Instruct-ERIC. Access is granted on approval of the proposal via peer review.

“Analysed Data”: All data resulting from the manual or automated evaluation of Raw Data and Metadata through analytical and logical reasoning.

“Establishment”: The User’s employer.

“Experimental Data”: Raw Data, Analysed Data and associated Metadata arising from use of Instruct-ERIC Centre facilities

“Instruct-ERIC”:a structural biology distributed infrastructure and member of the ESFRI Roadmap.

"Instruct-ERIC Centre": an Institution recognized as a Centre by Instruct-ERIC providing Users with access to its experimental facilities, scientific skills and/or online resources in the context of an Access project

"Instruct-ERIC facilities": All facilities made available at Instruct-ERIC Centres

"Metadata": Information pertaining to Experimental or Analysed Data collected as a result of use of Instruct-ERIC Centre facilities and shall include (but shall not be limited to) the context of the experiment, the experimental team, experimental conditions and other logistical information.

"Proprietary Research": Commercially confidential research using Instruct-ERIC Centre facilities and for which there is no obligation to publish the Results.

"Raw Data": Data produced as a result of use of Instruct-ERIC Centre facilities, excluding Analysed Data and Metadata.

"Results": Any inventions, designs, information, know-how, specifications, formulae, Experimental Data, processes, methods, techniques and other technology arising out of peer reviewed research or activities.

"Structural data": We refer to experimentally-derived data such as, structure factors, structural maps, list of atomic coordinates, or information on interacting protein residues and interatomic distances, among many others;

"Supporting data": We refer to the data necessary to reproduce the published conclusions, including but not limited to original electron micrographs or particle images entering in the 3D reconstruction process, raw (time-domain) or processed (frequency-domain) NMR spectral data, diffraction data or other data arising from the use of X-ray sources.

"Users": Users shall include the following persons making use of Instruct-ERIC Centres through Instruct-ERIC access procedures: scientists and engineers from academia, research councils and charitable institutions, researchers from commercial and non-commercial organisations.

5. Data to which this policy applies

This policy applies to Experimental Data, Supporting Data and Structural data.

6. Data ownership

6.1 Centres will not claim any usage or IP rights on the Experimental Data that they produce.

6.2 Subject to pre-existing obligations including to various establishments, grant funding agencies or other third parties, and as a general rule, the Institution(s) to which the user belong while conducting Instruct-ERIC Access projects will have the exclusive use of the data acquired in the course of the project during the embargo period Indicated in the section 6. Intellectual Property protection for these Institutions and users will be their sole responsibility.

6.3 If the Experimental Data reveals problems or flaws in the technology used to acquire it, in the data processing procedures, or indicates that improved service and quality of service could be obtained, then, subject to agreement with the Instruct-ERIC centre user, the data may be used for the sole purpose of correcting these problems or flaws, or to improve the service.

6.4 In those cases in which the delivery of User Access requires a significant change or adaptation of otherwise standard Facility procedures, demanding substantial involvement of Facility staff, data ownership and IP will be shared between Users' Institutions and Facility. The details of this sharing

should be discussed at the time the modified Access protocols were designed, and the Facility will present the User's Institution a concrete proposal for discussion.

7. Data archiving

7.1 Subject to the pre-existing obligations above, storage of data is the responsibility of the User/Institution to whom it belongs. Unless the Instruct-ERIC Centre explicitly offers a data archive service, Users are responsible for copying and making arrangements for the long term storage of the Experimental data. In this latter case the facility will collect and maintain an accountable proof of the transfer of the data to the user, for verification purposes.

7.2 Subject to future developments in the context of the European Open Science Cloud (EOSC) and of the approval of the appropriate Instruct-ERIC internal protocols of actions, and if in the future the EOSC initiative will provide the opportunity to store all, or a subset, of the experimental data acquired at Instruct-ERIC Centres in the course of Instruct-ERIC Access projects, then Instruct-ERIC will implement the necessary actions to take full advantage of such an opportunity. These data will be appropriately labelled using, for instance, Digital Object Identifiers (DOI).

8. Data sharing

8.1 Structural data and models obtained in the course of the research conducted within an Instruct-ERIC Access project must be deposited in an appropriate public database. In particular, structural data must be either deposited in PDB/EMDB or, as an exception, to be made otherwise available within one year after publication of the results, or within five years after the visit, whichever came first.

8.2 It is the responsibility of the user to assure that supporting data is deposited in a public database or, in the absence of an appropriate such database, made otherwise available within one year after publication of the results, or within five years after the visit, whichever came first.

9. Data confidentiality

Instruct-ERIC, and Instruct-ERIC Centres, shall have procedures/guidelines in place to ensure confidentiality, both internally & externally, of Experimental Data during the embargo period, as well as to use these procedures to ensure that access to Experimental Data will be restricted to the Users of Instruct-ERIC Access projects to which the Experimental Data relates and the appropriate Instruct-ERIC support staff. Users of Instruct-ERIC facilities are responsible for meeting any third party data management or transfer obligations that may be applicable.

B.6. LAGO

Version	Date	Contributors
1.0	17/2/2020	Antonio Juan Rubio Montero (CIEMAT), Rafael Mayo (CIEMAT)

B.6.1. Data summary

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation.** The Latin American Giant Observatory (LAGO) is an extended cosmic ray observatory composed of a network water-Cherenkov detectors (WCD) spanning over different sites located at significantly different altitudes and latitudes. The measurements collected from these detectors are posteriorly processed and analysed. Additionally, scientists continuously generate simulated data. The final purpose is to enable the long-term curation and re-use of data within and outside the LAGO Collaboration through a Virtual Observatory.
- **Explain the relation to the objectives of the project.** European Commision requires open access to the results obtained from their funded projects, meanwhile EOSC-Synergy is a H2020 project that encourages the implementation of FAIR policies as another standard in research. Since the LAGO computation is included in the EOSC-Synergy as one of their Thematic Services, generated or stored data in its resources must observe these guidelines, being also beneficial for the success of both initiatives.
- **Specify the types and formats of data generated/collected.** CORSIKA outputs described in the official documentation [D. Heck and T. Piero, "Extensive Air Shower Simulation with CORSIKA: A User's Guide". Version 7.7100 from December 17, 2019], section 10, page 121. Available at https://web.ikp.kit.edu/corsika/usersguide/usersguide.pdf
- **Specify if existing data is being re-used (if any).** Measurements from WCDs gathered in previous years.
- **Specify the origin of the data.**
    - Raw data (L0) from WCDs .
    - Preliminary data (L1) obtained cleaning raw data (L0)
    - Quality data (L2, L3) obtained analysing and fixing preliminary data (L1).
    - Simulated from standalone CORSIKA runs by researchers.
- **State the expected size of the data (if known).** Minimal data-set is one hour of measurement or simulation:
    - Raw data (L0): ~200MB
    - Preliminary data (L1): ~100MB
    - Quality data (L2, L3): ~ 30 MB
    - Simulated (background): ~ 10GB
    - Simulated (event): ~ 100GB

- **Outline the data utility: to whom will it be useful.** Data are of interest for the Astrophysics community but also for other scientific or industrial areas such as High Energy Physics, Life Sciences, Weather Forecasting, Aerospatial security or Computer Science, among others, because the effects of cosmic radiation on natural life, materials, or climate change.

B.6.2. FAIR data

B.6.2.1 Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision)**. Specific LAGO wrappers execute the processing or simulation and posteriorly check the data-sets and will store them in EGI DataHub always with their metadata to allow gathering by services such as B2FIND.
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?** Data-sets will be referenced by PIDs automatically requested through EOSC B2Handle service.
- **Outline naming conventions used.** It should be based in the metadata values but an approach for clear versioning is being discussed.
- **Outline the approach towards search keywords.** Searching should be based on any metadata value.
- **Outline the approach for clear versioning.** It should be based on the metadata An approach for clear versioning is being discussed.
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**

B.6.2.2 Making data openly accessible:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so.** Data will be made publicly available after a variable waiting (embargo) period similar to the established ones for other large experiments.
- **Specify how the data will be made available.** Consolidated data-sets that are stored in EGI DataHub will be exposed together with their metadata to be gathered by services such as B2FIND.
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?**. To take advantage of the data published, researchers should use the CORSIKA tools included in the source code and described in the official documentation in section 10, page 121 at https://web.ikp.kit.edu/corsika/usersguide/usersguide.pdf
- **Specify where the data and associated metadata, documentation and code are deposited.**
    - Data and metadata will be stored in EGI DataHub service (OneData technology)
    - CORSIKA documentation and source code https://web.ikp.kit.edu/corsika/

- **Specify how access will be provided in case there are any restrictions.** Data will be only accessible by the author and/or the Collaboration making during embargo period use of EGI AAI.

B.6.2.3 Making data interoperable:

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.** Metadata follows the Dublin Core schema (http://dublincore.org), extending the vocabulary with the elements the described in [H. Asorey et al. The LAGO: A Successful Collaboration in Latin America Based on Cosmic Rays and Computer Science Domains, in Proc. 16th IEEE/ACM CCGrid, 2016, https://doi.org/10.1109/CCGrid.2016.110].
    - Common for all metadata: *site* contains the *name, latitude, longitude* and *height* of the WCD or the simulated ground point.
    - WCD metadata scheme adds: *data* corresponds to the *version/type* of the Digit/Analog electronic board; *voltage, level* and *sensor*.
    - Simulation metadata adds: *primary* described by the CORSIKA input file DATXXXX.dbase; *libraries* indicating which are the included CORSIKA libraries; *computation* describing the computational environment by unix command: *uname -a*, *lsb_release -a*, *free* and *gcc -v*.
- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?** In principle, only support CORSIKA outputs as described in the official documentation, but we can consider translating files to standardised formats in the future.

B.6.2.4 Increase data re-use (through clarifying licenses):

- **Specify how the data will be licenced to permit the widest reuse possible.** They will be published under BSD-3 or CC license.
- **Specify when the data will be made available for reuse. If applicable, specify why and for what period a data embargo is needed.** LAGO Collaboration requires a waiting period similar to the established ones for other large experiments. Such a period should be set not only to properly exploit results by the Consortium prior to their availability, but because raw data measured must be pre-processed by the Consortium to make them 'understandable'. Simulations will be available too, but it would be valuable that the waiting period could be set by the user, because he is the owner of the data. The embargo period is set for a year in general, but depends of the data type, specifically:
    - L0, L1: private while analysed data are not publicly available.
    - L2, L3: a year.
    - Simulated data: a year maximum, the owner can decide to open the data before the end of this period.
- **Specify whether the data produced and/or used in the project is usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.** There is no restriction after the embargo period.

- **Describe data quality assurance processes.** Only the data measured by WCDs or generated using software versions officially released by LAGO will be stored and exposed in repositories. Previously to the publication, a robot of the Virtual Organization will check the minimal accuracy of data.
- **Specify the length of time for which the data will remain re-usable.** Indefinitely after the waiting period.

B.6.3. Allocation of resources

Explain the allocation of resources, addressing the following issues:

- **Estimate the costs for making your data FAIR. Describe how you intend to cover these costs.** The process of making the data FAIR will be supported by the EOSC-Synergy project. The human cost of the management will be supported by LAGO Collaboration.
- **Clearly identify responsibilities for data management in your project.** Computing as data management will be structured as a Virtual Organization with specific roles for data acquisition and processing.
- **Describe costs and potential value of long term preservation.** Preservation of data-sets is essential for the sustainability of LAGO. Every active WCD should generate 300GB/month of L0-L3 data. Currently, due to the number of active WCDs, the Collaboration will generate up to 27 TB of L0-L3 data, plus 12-120 TB of simulated data throughout the year. Data should be replicated, at least, in two locations of a distributed repository (in this case OneData). Considering an average generation of 60TB/year, the costs of long-term preservation for 4 years are the hardware (two generic RAID servers ~240TB = ~30k€, prices in 2019), the consumption (3.68KW max. power for 2 servers, ~ 0.1 €/kWh industrial price average in 2019 = max. 13k€) and human resources (technician: 1 person/month, scientific: 2 p/m, ~10k€).

B.6.4. Data security

Address data recovery as well as secure storage and transfer of sensitive data. There is no sensitive data, thus anonymization and encryption of the data is not required. Data recovery should be guaranteed by means of replication, at least, in two locations of a distributed repository or filesystem (in this case OneData).

B.6.5. Ethical aspects

Data do not contain protected records that could present ethical or security issues. The only personal data included is the required by FAIR policies in metadata, this is, the name and identifier of the author of the data-set. On the other hand, there are no issues with reusing previous raw data generated in LAGO, as well as the data belonging to the Collaboration.

B.7. SDS-WAS

Version	Date	Contributors
1.0	18/2/2020	Francesco Benincasa (BSC)

B.7.1. Data summary

- **State the purpose of the data collection/generation.** SDS-WAS data is stored in an in-house shared storage file-system. Data can be classified in two types, models outputs and observations. Models outputs consist in a set of 12 NWP (Numerical Weather Prediction) model outputs of two variables (dust surface concentration at sea level and aerosol optical depth of the whole column) with 72 hours forecast (3/6 hourly) at various spatial resolutions from 0.33º to 0.5º approximately. Of these models 2 are run in house in an HPC infrastructure and the remaining are collected from partner institutions with a variety of protocols/methods: http, ftp, receiving, downloading, etc. This data is disseminated by the thematic service in a documented standard format. On the other hand observations are collected from a variety of sources to perform model evaluation and validate results a set of observations is downloaded. They are not disseminated because they are publicly available from their respective official sources (mostly NASA currently). Models outputs are processed to a common data standard following netCDF format and CF-1.6 conventions. Observations come in different formats, which are processed and formatted to be compared with model data. Furthermore, a wide range of derived products are derived from data analysis, in numerical (netCDF) format, in picture formats (png, animated gifs) and text table format (numerical scores).
- **Explain the relation to the objectives of the project.** The thematic service SDS-WAS through the integration in EOSC aims to reach more potential users also outside the scientific domain specific community, improve data FAIRness and synergies with other EOSC services.
- **Specify the types and formats of data generated/collected.** The data format used shall be Network Common Data Form (netCDF)[21]. Data in netCDF format is self-Describing, portable, scalable, appendable, shareable, and archivable. The metadata used shall follow the conventions for CF (Climate and Forecast) metadata[22].
- **Specify if existing data is being re-used**. The service uses data generated by itself and from other sources.
- **Specify the origin of the data.** Observations data is referenced to the official sources. Models data is referenced to respective owners/developers, and derived products workflow is well documented.
- **State the expected size of the data.** Data growing is related to some factors like the increase of models number joining to the project, the increase of observations collected, the derived products generated and the resolution (temporal and spatial) of all previous described data. By now the occupied storage is about 4TB and the tendency is to increase ~1TB per year.

[21] https://www.unidata.ucar.edu/software/netcdf/docs/netcdf_introduction.html
[22] http://cfconventions.org/

- **Outline the data utility: to whom will it be useful.** Data collected and disseminated is very useful for all researchers working on Sand and Dust Storms field, plus all related implications (health, industry, etc ...).

B.7.2. FAIR Data

B.7.2.1 Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision).** Data shall be published using the netCDF[23] data format with metadata following the CF convention[24].
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?** We still do have neither PIDs nor DOIs but we plan to introduce such a kind of service.
- **Outline naming conventions used.** Naming convention follows the CF Standard Names[25].
- **Outline the approach towards search keywords.** An approach towards search keywords is being discussed.
- **Outline the approach for clear versioning.** An approach for clear versioning is being discussed.
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how.** Data shall be published using the netCDF data format with metadata following the CF convention.

B.7.2.2 Making data openly accessible:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so.** Observational data are not disseminated and referred to respective external sources. Model data are disseminated according to the following policy:
    - Data can be downloaded through free portal login for tracking purposes.
    - Data with 2 days delay have freely available to all registered users.
    - Near Real-Time (NRT) data are freely available only to partners and in general non-commercial users who explicitly ask for it.
    - Post-processed products images are freely available without registering.
- **Specify how the data will be made available.** Processed data shall be available via web interface, programmes via git.
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?** Data can easily be accessed by freely and openly available netCDF data viewers[26].
- **Specify where the data and associated metadata, documentation and code are deposited.** Long term storage shall be on B2SAFE local instance

[23] https://www.unidata.ucar.edu/software/netcdf/docs/netcdf_introduction.html
[24] http://cfconvention.org
[25] http://cfconventions.org/Data/cf-standard-names/current/build/cf-standard-name-table.html
[26] https://www.unidata.ucar.edu/software/netcdf/software.html

- **Specify how access will be provided in case there are any restrictions.** To be discussed

B.7.2.3 Making data interoperable:

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.** Data will be interoperable through use of the netCDF data format with metadata following the CF convention.
- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?** Standard vocabulary is described using the netCDF data format with metadata following the CF convention.

B.7.2.4 Increase data re-use (through clarifying licenses):

- **Specify how the data will be licenced to permit the widest reuse possible.** Ongoing discussion.
- **Specify when the data will be made available for reuse. If applicable, specify why and for what period a data embargo is needed.** Data can be used immediately after publication with the previous mentioned restrictions (NRT only for project partners and non-commercial users who explicitly request them, and 2 days delay data available for everyone through registration to the web portal).
- **Specify whether the data produced and/or used in the project is usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.** Data can be reused.
- **Describe data quality assurance processes.** Data quality of the source data is assured by the data providers. Processed data will be secured on servers without write access to the general public, in order to prevent data manipulation.
- **Specify the length of time for which the data will remain re-usable.** To be discussed.

B.7.3. Allocation of resources

Data production and storage is financed by SDS-WAS consortium made by AEMET (Spanish Meteorological Agency) and Barcelona Supercomputing Center (BSC). Data management is in charge of BSC. Long preservation is still under discussion.

B.7.4. Data security

Data is replicated from a safe data archive. The replicated data will be backed-up.

B.7.5. Ethical aspects

No ethical issues in terms of data generation or usage exist

B.8. UMSA

Version	Date	Contributors
1.0	17/2/2020	Aleš Křenek (CESNET)

B.8.1. Data summary

- **State the purpose of the data collection/generation.** The datasets are acquired and processed as a part of various studies of "exposome", i.e. research of joint effects of exposure of humans to various factors (environment, health, ...).
- **Explain the relation to the objectives of the project.** The research spans multiple groups in Europe and worldwide, the studies aim at gathering data for long term typically, EOSC is expected to become the standard platform for data sharing and processing in this community.
- **Specify the types and formats of data generated/collected.** Vendor proprietary raw formats of mass-spectrometry profiles, the profiles converted to open formats (mzML will prevail), and associated metadata in json.
- **Specify if existing data is being re-used (if any).** No, only data acquired during the project.
- **Specify the origin of the data.** Mass spectrometers in the laboratories of the users.
- **State the expected size of the data (if known).** Will grow gradually up to hundreds of Terabytes in a few years.
- **Outline the data utility: to whom will it be useful.** The exposome research community (emerging EIRENE ESFRI). Most of the data are part of longitudinal studies, they will be revisited for decades.

B.8.2. FAIR data

B.8.2.1 Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision.** The service will be interfaced to other services of the community, which will implement the discoverability. This happens at two levels:
    - Each dataset is uniquely assigned to a sample (and the samples are linked to specific studies, patients, environment measurements etc.); the datasets are discoverable by the sample ID.
    - The datasets contain derived data (typically, compounds or their properties identified in the samples) that are fed into indexing engines, and are searchable by those properties.
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?.** Internal persistent and immutable identifiers are applied to all principal entities in the whole system (studies, people involved, samples, as well as the datasets). Standardized identifiers (DOI) will be assigned to the "top level" entities only (studies) to keep their number manageable.

- **Outline naming conventions used.** Internally, datasets are named hierarchically (year, study, batch, sample, ...), however, the convention is mostly irrelevant for the view from outside. We rely on findability described above.
- **Outline the approach towards search keywords.** Simple keyword search is applicable at the topmost (study) level only; it is irrelevant elsewhere.
- **Outline the approach for clear versioning.** Primary datafiles are unique, the sample is physically destroyed while acquiring the data. Derived data are versioned thoroughly, always keeping the information of the workflow (including versions of the included individual tools).
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how.** The metadata schema will emerge over time, there are no current standards. In general, the metadata describe the laboratory processes used to acquire the data.

B.8.2.2 *Making data openly accessible:*

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so.** The general community policy is making as much openly available as possible. However, the data come from studies which may enforce different policies (in particular, when originating from human samples).
- **Specify how the data will be made available.** Download from the website, after the dataset was identified.
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?** Primary experimental data will be available in the open mzML standardized format, accepted by virtually all software in the application domain. Processed data are simple tables in CSV typically.
- **Specify where the data and associated metadata, documentation and code are deposited.** Metadata related to the primary experimental data are always stored with the data, in machine and human readable format (json). Search over the metadata will be available by other services being setup in EIRENE.
- **Specify how access will be provided in case there are any restrictions.** The processes are still to be defined, based on experience with emerging use cases.

B.8.2.3 Making data interoperable:

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.** Untargetted mass spectrometry is in a premature stage from this viewpoint, no standards exist yet. Though the standards will emerge in approx. 5 years time frame.
- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?** Not applicable.

B.8.2.4 Increase data re-use (through clarifying licenses):

- **Specify how the data will be licenced to permit the widest reuse possible.** It will depend on specific restrictions of studies that provide the data. In general, the community agrees on as much open approach as possible, therefore Creative Commons or similar license families are expected.
- **Specify when the data will be made available for reuse. If applicable, specify why and for what period a data embargo is needed.** Typically, an embargo will be held till publication of the related papers only.
- **Specify whether the data produced and/or used in the project is usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.** The services are part of emerging infrastructure, its primary purpose is making the data available for long term.
- **Describe data quality assurance processes.** Acquisition of the primary data follows the quality control procedures of the providing laboratories. Derived data are generated with user workflows, with no specific quality control enforced; however, because of strict provenance checking, it is always possible to trace down the process leading to a particular data set.
- **Specify the length of time for which the data will remain re-usable.** Decades are expected.

B.8.3. Allocation of resources

Making data FAIR is intrinsic in this infrastructure, it is its principal purpose. Therefore estimation of the "added FAIRness" cost is pointless. Similarly, the purpose of the infrastructure is making the data available for long term to enable longitudinal studies. The overall long term cost evaluation is subject of work of the running projects which prepare setup of the ESFRI EIRENE infrastructure.

B.8.4. Data security

Data are stored at CESNET and Masaryk University data storage, using CESNET backup services (two copies in different locations). Sensitive personal information is not included in the data. Pieces of the data that are not publicly available will be protected by access allowed to authenticated users only, and standard access control mechanisms of the underlying storage systems.

B.8.5. Ethical aspects

Ethical aspects, if any, are covered by specific scientific studies that use the service.

B.9. MSWSS

Version	Date	Contributors
1.0	17/2/2020	Jan Astalos (IISAS)

B.9.1. Data summary

- **The purpose of the data collection/generation:** MSWSS is a service for analysis of water distribution network with regards to the mitigation of hazardous events by the integration of existing on-line analysis of toxics in drinking water supply networks with water distribution network simulation (EPANET). Other potential uses of the service are rehabilitation planning and optimisation. Analysis of hazardous events may be used for preparation of risk management plans for water utilities with potential to be extended to an on-line early warning system. In addition to the use by water infrastructure operators the service may be used also for research and educational purposes.
- **Relation to the objectives of the project**: MSWSS is one of the Thematic Services of the EOSC-Synergy project. The aim of the project is to promote EOSC adoption by the research communities, represented by the Tematic services, by expanding and building knowledge on common interfaces, standards and best practices. This will be supported by an expansion of the capacity through the federation of resources aligned with the EOSC and FAIR policies and practices.
- **Data reusability:** Operational data may have confidential status. In this case the data will be re-used only by the user who owns them (who uploaded them to the MSWSS service or who obtained them as an output from processing in MSWSS service). If the service will be used for research, the data re-usability will be under control of the data owner.
- **The origin of the data:** Some of the data originate from the users (Water Supply System operators or researchers) themselves (GIS, CIS and SCADA data) and some are publicly available data (data supporting pre-processing and/or post-processing = ZBGIS, OpenStreetMap, DEM50 data).
- **The expected size of the input data:** The size of input data for each job is approximately 800 MB. The expected size of output data per job is approximately 200 MB. The size of the data will depend on the analysed water distribution network.
- **The data utility:** The output data from simulations and post-processing tasks performed in MSWSS service will be mainly used for further processing by users who performed the simulations (WSS operators or researchers).

B.9.2. FAIR data

B.9.2.1 Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision):** MSWSS service will not provide publicly accessible persistent storage for the data. The data stored in MSWSS service will remain private to

their respective users. If users decide to publish the data, they will have to store the data into some of the EOSC data repositories (e.g. ZENODO) and register the datasets to appropriate metadata search engines. The data will be then discoverable by standard means (e.g. keyword search).

- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?** MSWSS service will not manage persistent identifiers such as DOI. The users will have to obtain the identifiers from external sources.
- **Outline naming conventions used:** The naming convention is yet to be defined.
- **Outline the approach towards search keyword:** The keywords for the data will be provided by users.
- **Outline the approach for clear versioning:** The data versioning is being discussed.
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how.:** The standard for metadata creation is being discussed.

B.9.2.2 Making data openly accessible:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so:** The data used for research or educational purposes could be made openly available by their owners. The operational data may be confidential by national legislation and/or institutional policies.
- **Specify how the data will be made available:** The users will have to store the data into some of the publicly available EOSC data repositories (e.g. ZENODO).
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?** The protocols for accessing the data will depend on the selected data repository.
- **Specify where the data and associated metadata, documentation and code are deposited:** This is yet to be defined.
- **Specify how access will be provided in case there are any restrictions.** This will depend on the selected repository and its access control mechanisms.

B.9.2.3 Making data interoperable:

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.** This is yet to be defined.
- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?** This is yet to be defined.

B.9.2.4 Increase data re-use (through clarifying licenses):

- **Specify how the data will be licenced to permit the widest reuse possible:** Research or educational data are expected to be open. The license will be selected by data owners. Access to the operational data may be limited by national laws or institutional policies.
- **Specify when the data will be made available for reuse. If applicable, specify why and for what period a data embargo is needed:** No embargo is currently foreseen, however, it will depend on the data owners.
- **Specify whether the data produced and/or used in the project is usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why:** Research data could potentially be re-usable by third parties.
- **Describe data quality assurance processes:** No automatic data quality assurance processes will be implemented in MSWSS service.
- **Specify the length of time for which the data will remain re-usable:** This will depend on the data owners.

B.9.3. Allocation of resources

- **Costs for making your data FAIR:** Making data FAIR will require costs associated with the development of necessary functionality into MSWSS service. However, as the external EOSC data repositories (e.g. ZENODO) are planned be used for storing the data, the additional functionality will be minimised. If the users will request a dedicated data repository, this plan will be revised and the associated costs estimated.
- **Responsibilities for data management:** The responsibility for data management will be on the owners of the data.
- **Costs and potential value of long term preservation:** Long term preservation of the data is not currently planned.

B.9.4. Data security

Because of the data confidentiality requirements for operational data the access to the data will be restricted to their respective users. The MSWSS service will provide means for ensuring data confidentiality at all levels of processing. Data security at the level of IaaS will be negotiated individually with the resource providers and will be included in the Service Level Agreements. The data will be transferred through encrypted connections.
If the service will be used for research or educational purposes (i.e. no confidential data will be used), standard EOSC policies/procedures for data security will be sufficient.

B.9.5. Ethical aspects

No ethical aspects related to the data are envisaged.

B.9.6. Other

The access to operational data may be regulated by national legislation (e.g. in Slovak Republic “Critical infrastructure law (No. 45/2011)”) and further by the institutional/departmental procedures or policies of the respective users.

B.10. O3AS

Version	Date	Contributors
1.0	17/2/2020	Tobias Kerzenmacher (KIT)

B.10.1. Data summary

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation.** Two different types of data have to be distinguished, firstly input data needed to provide the O3AS thematic service, and secondly data that is generated and provided through the O3AS service. Input data from the IGAC/SPARC Chemistry Climate Model Initiative[27] provided by the CEDA Archive[28] and possibly ERA-Interim[29] or ERA5[30]. These data sets are used to produce trend analyses of ozone which shall be made available by the O3AS thematic service to scientists and educational institutions as plots and ozone time series for the reoccuring ozone assessment analyses and for educational purposes (atmospheric ozone chemistry, atmospheric circulation) in schools and universities.
- **Explain the relation to the objectives of the project.** The thematic service O3AS will be made available to international scientists thereby promoting the EOSC service. It builds on the existing knowledge of already present standards in the climate sciences making it available and promoting it to other services, encouraging the discussion of common interfaces, standards & best practices.
- **Specify the types and formats of data generated/collected.** The data format used shall be Network Common Data Form (netCDF)[31]. Data in netCDF format is self-Describing, portable, scalable, appendable, shareable, and archivable. The metadata used shall follow the conventions for CF (Climate and Forecast) metadata[32].
- **Specify if existing data is being re-used.** Input data from the IGAC/SPARC Chemistry Climate Model Initiative provided by the CEDA Archive and possibly ERA-Interim or ERA5 are solely used for producing the trend data sets offered by the O3AS thematic service.
- **Specify the origin of the data.**Input data is available upon registration[33] from the CEDA Archive and ECMWF[34] or the Climate Data Store[35]. The data access policy for the CEDA ArchiveAfter processing this data sets we offer ozone time series through the O3AS thematic service.
- **State the expected size of the data.** An example of model data for 40 years (512 longitudes, 256 latitudes and 37 heights with about 60000 time steps) showed a data reduction from 4TiB to 60

[27] http://blogs.reading.ac.uk/ccmi/
[28] http://archive.ceda.ac.uk
[29] https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim
[30] https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5
[31] https://www.unidata.ucar.edu/software/netcdf/docs/netcdf_introduction.html
[32] http://cfconventions.org/
[33] https://blogs.reading.ac.uk/ccmi/badc-data-access/
[34] https://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/
[35] https://cds.climate.copernicus.eu/#!/search?text=ERA5&type=dataset

kiB for ozone. The time series of the data offered shall be longer by a factor of three to four. Also the service shall be offered for about 10 models which results in an expected size of 200 TiB of primary data for which only temporary storage is required and which will be reduced to more permanent 3 MiB that shall be offered at the the Large Scale Data Facility (LSDF) at KIT through the O3AS thematic service. Further refinement of the processing workflow (e.g. choice of month to plot, integrate ozone over height, smoothing).

- **Outline the data utility: to whom will it be useful.** The data will be of invaluable use for the scientists working on the Ozone Assessment report[36]. Additionally, it will provide a tool for educational institutions illustrating the problem of the ozone whole and how it is being managed.

B.10.2. FAIR data

B.10.2.1 Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision).** Data shall be published using the netCDF data format with metadata following the CF convention.
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as DOIs?** Data published at the Karlsruhe Institute of Technology[37] or within the Helmholtz Data Federation[38] shall be provided with a DOI.
- **Outline naming conventions used.** Naming convention follows the CF Standard Names of CF convention.
- **Outline the approach towards search keywords.** An approach towards search keywords is being discussed.
- **Outline the approach for clear versioning.** An approach for clear versioning is being discussed.
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how.** Data shall be published using the netCDF data format with metadata following the CF convention.

B.10.2.2 Making data openly accessible:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so.** All processed data (ozone trend time series) and the programmes for processing the data shall be openly available under an apache-v2, an MIT or a GPL license.
- **Specify how the data will be made available.** Processed data shall be available via web interface, programmes via git.
- **Specify what methods or sw tools are needed to access the data? Is documentation about the sw needed to access the data included?** Data can easily be accessed by freely and openly available netCDF data viewers[39].

[36] https://www.esrl.noaa.gov/csd/assessments/ozone/
[37] http://www.kit.edu/english/
[38] https://www.helmholtz.de/en/research/information-data-science/helmholtz-data-federation-hdf/
[39] https://www.unidata.ucar.edu/software/netcdf/software.html

- **Specify where the data and associated metadata, documentation and code are deposited.** Long term storage shall be on LSDF and RADAR4KIT (Research Data Repository in development).
- **Specify how access will be provided in case there are any restrictions.** It is foreseen that access will be provided openly.

B.10.2.3 Making data interoperable:

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.** Data will be interoperable through use of the netCDF data format with metadata following the CF convention.
- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability?** Standard vocabulary is described using the netCDF data format with metadata following the CF convention.

B.10.2.4 Increase data re-use (through clarifying licenses):

- **Specify how the data will be licenced to permit the widest reuse possible.** Whenever possible the GNU General Public License v3.0 shall be used.
- **Specify when the data will be made available for reuse. If applicable, specify why and for what period a data embargo is needed.** Data can be used immediately after publication.
- **Specify whether the data produced and/or used in the project is usable by third parties, in particular after the end of the project?.** Data can be reused.
- **Describe data quality assurance processes.** Data quality of the source data is assured by the data providers. Processed data will be secured on servers without write access to the general public, in order to prevent data manipulation.
- **Specify the length of time for which the data will remain re-usable.** The processed data sets shall remain available for 20 years.

B.10.3. Allocation of resources

Explain the allocation of resources, addressing the following issues:

- Estimate the costs for making your data FAIR. Describe how you intend to cover these costs.
- Clearly identify responsibilities for data management in your project.
- Describe costs and potential value of long term preservation

Data production is largely project based – the data products are close to FAIR; long-term data storage is financed by strategic investment. National differences exist. Climate research and the required infrastructures are funded by the Helmholtz Programme. This ensures the required long term funding required, unless otherwise decided by the German Ministry for Science and Education.

B.10.4. Data security

Data is replicated from a safe data archive. The replicated data will be backed-up.

B.10.5. Ethical aspects

No ethical issues in terms of data generation or usage exist.

B.10.6. Other

At the Karlsruhe Institute of Technology scientists are supported through a research data management team (RDM@KIT).